scientific reports

OPEN

Creating the Slovenian genome database and browser as a source of comprehensive variation of the Slovenian population

Aleš Maver[1,2], Peter Juvan[1], Urška Kotnik[1], Luca Lovrecic[1,2], Gaber Bergant[1] & Borut Peterlin[1,2]✉

The genomic data of Central European populations is underrepresented in the publicly available databases. We present the comprehensive genomic variation of the Slovenian population, based on the genomic sequencing of 9425 non-related individuals, i.e. more than 0.44% of the Slovenian population. Over 30 million unique single nucleotide and small indel (30.8 million), copy number (217.6 thousand), and mitochondrial variants (3.3 thousand) were uncovered and annotated by analysing the whole genome of 619 individuals and the whole exome of 8806 individuals. This population variation, including 3,9 million novel variants, is presented in a publicly available genome variant browser, the SloGenVar (https://slogenvar.si). We used this newly developed resource to reveal the population frequency of pathogenic variants in the genes associated with recessive conditions. The Slovenian genome database and browser offer the largest and the most comprehensive publicly available Central European population genomic variant resource, providing an important asset for genomic studies and as a control variant database for variant interpretation in the region and beyond.

Keywords Population genomic variation, Whole genome sequencing, Genome browser, Slovenian genome database

Genomic variation is responsible for the shaping of population diversity and can reveal a variety of different traits in human populations, including various physical and medically significant traits such as disease susceptibility and drug responses. These traits can arise from rare variants, common variants, and various combinations thereof[1,2]. Thus, the study of population genomic variation is necessary for advancements in understanding human population structure[3] and the genomic basis of complex and rare diseases[4,5].

The increased availability of genomic sequencing and large-scale genome variation databases marks an important milestone in the study of genomic variation. As next-generation sequencing has continued to grow in demand and decline in cost over the last decade, populational databases have provided us with an unprecedented amount of information, therefore presenting an opportunity for improving our understanding of the diversity of the human genome[6]. Since their creation, many international large-scale genome variation databases have become an integral part of genomic research[7–9], providing us with easily accessible population-based information, such as data on both rare and common variant frequencies and common pathogenic variants within the investigated populations[10]. This has enabled an unprecedented opportunity for studying the role of rare and common sequence variants in biological processes, complex traits, and rare diseases[9,11].

Previous big sequencing projects have described more than 99% of common variants present in the human population, providing both a reference of common variants on a global scale and an assessment of their distribution, delivering a basis for further population genomics research[7,9,12,13]. Nevertheless, despite extensive work in capturing human genomic variation, not all the world's regions are represented equally. The world's most extensive genomic database, gnomAD, contains more than 800,000 control individuals from different populations, yet approximately half of this cohort comes from the UK biobank[9]. Other populations, including other non-Western European populations, only present a minority in this and other published genome variation databases[14,15]. Importantly, rare genomic variants are often unique to specific populations and present a frequent cause of monogenic disorders[16,17], and the genomic data varies between populations, even among geographically closely related European populations[18]. Consequently, comparing the genomic variation from different

[1]Clinical Institute of Genomic Medicine, University Medical Centre Ljubljana, Ljubljana, Slovenia. [2]Present address: Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia. ✉email: borut.peterlin@kclj.si

populations provides insight into the genomic diversity, relation amongst different populations, and the human population history[19,20] as well as identifying unique characteristics of a specific population, including founder variants and population-specific frequency of pathogenic variants[21]. This data supports the development of precision medicine-based health strategies, such as population-specific screening protocols, targeted prevention approaches, informed education and research, and risk-based development of national health infrastructure for disease prevention, diagnosis, monitoring, and treatment, with realistic expectations regarding the disease risks and treatment benefits in the population[22,23].

National sequencing projects have been an important area of research in the last decade[24], with an increasing number of countries and populations added to the list every year. Many of the world populations have so far initiated their large-scale genomic projects and published accounts of genomic variation within populations, representing various regions from Africa[25], Europe[26], America[27], Asia[28–30], and Australia[31]. The objectives of those population/national genome projects are to characterize population-specific rare disease variants, improve genomic diagnostics in analysed populations, and introduce public health genomics in public health policies. Moreover, the diversity and thus quality of published genomic data is improved by the inclusion of a variety of previously under-represented populations in openly available genomic databases. European populations outside of Western and Northern Europe have so far been scarcely included in such efforts[24], with the Czech Republic[32] and Poland[33] being the only notable exceptions in the Central European region with large-scale national sequencing projects, encompassing approximately 0.01% of their population[24]. A Slovenian genome project was initiated in 2018[34] and has presented a facilitator for the analysis of the Slovenian genome database, presented in this paper.

The present work aims to fill the gap in the representation of Central European populations in the publicly available large-scale genome variation databases. Here we present the genomic variation of the Slovenian population, which includes whole genome and exome sequences of 9,425 non-related individuals of Slovenian ancestry, presenting more than 0.44% of the Slovenian population. The Slovenian genome database contains over 30 million variants and is presented and publicly available in a user-friendly browser (https://slogenvar.si). We present the cumulative variability of the Slovenian population, encompassing both rare and common genomic variation, including single nucleotide variants, mitochondrial variants, copy number variation, and prevalence of pathogenic variants in genes connected to recessive disorders in the population.

This study provides one of the first characterizations of genomic population structure in the Central European region and one of the first descriptions of mitochondrial variants and copy number variants on a national scale. This initiative hopes to provide researchers worldwide with additional information on variant frequencies in the region, allowing for improved genetic data imputation and biomedical research as well as a control database for interpretation of genomic variants specifically in previously under-represented Central European region.

Results

Variation in the Slovenian population

Presented below is all the detectable high-quality genomic variation in the Slovenian genome database consisting of 9,425 non-related individuals of Slovenian origin, from the whole genome sequencing of 619 individuals and the whole exome sequencing of 8,806 individuals. The variation is composed of single nucleotide variants, mitochondrial variants, and copy number variants.

Single nucleotide variants

30,836,190 small variants, identified in our study, are composed of 26,674,174 single nucleotide variants and 4,162,016 indel variants (Table 1). This high-quality dataset presented the basis for the SNV distribution calculations. The distribution of indel variants is presented in Supplementary Fig. 1, revealing that the number of small indel variants decreases as their length increases, for both deletions/duplications.

		WGS + WES	WGS	WES
	Variant type	Number (%)	Number (%)	Number (%)
Non-coding 30,258,529 (98.13%)	Intron	19,493,709 (63.22)	19,305,670 (63.78)	188,039 (33.29)
	Intragenic	6,082,247 (19.72)	6,082,247 (20.09)	0 (0.00)
	Regulatory	4,238,374 (13.74)	4,180,393 (13.81)	57,981 (10.27)
	Other non-coding	444,199 (1.44)	440,431 (1.45)	3,768 (0.67)
Coding 577,661 (1.87%)	Missense	324,664 (1.05)	140,466 (0.46)	184,198 (32.61)
	Synonymous	204,020 (0.66)	99,018 (0.33)	105,002 (18.59)
	Canonical splice	18,691 (0.06)	10,781 (0.04)	7,910 (1.40)
	Frameshift	11,893 (0.04)	4,523 (0.01)	7,370 (1.30)
	In-frame indel	9,686 (0.03)	4,468 (0.01)	5,218 (0.92)
	Stop gain	8,562 (0.03)	3,364 (0.01)	5,198 (0.92)
	Other	145 (0.00)	60 (0.00)	85 (0.02)
	Sum	30,836,190 (100)	30,271,421 (100)	564,769 (100)

Table 1. Distribution of variation in the Slovenian population.

We have identified 30,258,529 (98.13%) non-coding variants (intronic, intragenic, regulatory, other) and 577,661 (1.87%) protein-coding changes (missense, synonymous, loss of function, in-frame indel, other). Most variants (98.17%) were detected in the WGS cohort.

Considering the WGS dataset only, the highest part of the identified SNV variants (99.13%) were in non-coding regions of the genome. Coding variants represent a minority, i.e. 0.87% of the found variants, where 0.46% were missense, 0.33% were synonymous, 0.06% were loss of function (canonical splice, frameshift, stop gain), 0.01% were in frame indel variants and the rest were other types of variants.

Amongst the variants, detected in the WES dataset, 44.23% of variants were non-coding, and the remaining 55.77% were coding variants: 32.61% were missense, 18.59% were synonyms, 3.63% were loss of function, 0.92% were in frame indel variants and the rest were other types of variants.

Overall, the results indicated that most of the variants in the Slovenian population were exceedingly rare, with the majority detected in frequency below 0.1%. We identified 3,896,636 variants (12,64%), not present in the gnomAD dataset (gnomAD 2.1), likely making them novel variation, specific to our dataset. Importantly, the vast majority of the novel variation in our cohort are rare variants (frequency < 0.1%) (Fig. 1a).

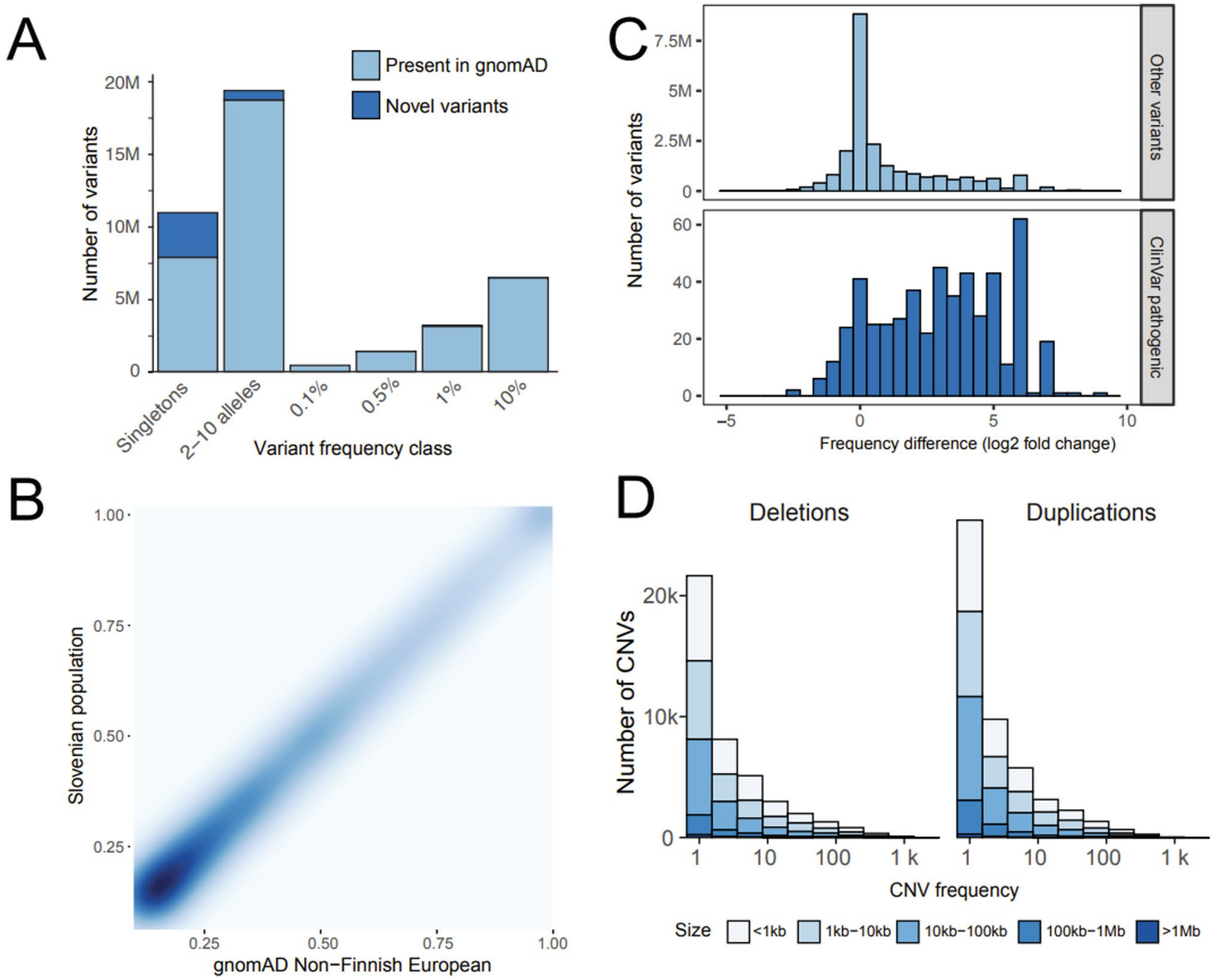

Fig. 1. Distribution of variant frequency of the variants detected in the Slovenian population. Comparison of variant frequency determined in the Slovenian population with gnomAD NFE population frequencies. (**a**) Novel variants are highlighted in dark coloring while existing gnomAD variation is represented in light coloring. (**b**) Variants exceeding the frequency of 10% in Slovenian and the selected populations are displayed. (**c**) Variant frequency differences of pathogenic variation in the Slovenian population compared with the gnomAD NFE frequencies. The ratio of frequencies in Slline numovenian versus gnomAD NFE is presented in log2 fold change values. The ClinVar pathogenic variants are presented in a separate panel, to illustrate a shift in frequencies of pathogenic variation. The results show an enrichment of rare variants in the Slovenian population as compared to the gnomAD dataset (the graph skewed in the positive direction) since the analyses included only variants, detected in the Slovenian population and gnomAD-only variants were excluded. (**d**) Histogram of CNV frequency. The deletions detected in the Slovenian population are presented on the left and the duplications are presented on the right. The CNV sizes are presented in different shades of blue.

When focusing on the variant level, we have explored differences in variant frequencies between our population and the gnomAD NFE population. Overall, this comparison indicated an anticipated consistency of frequency estimates between the datasets. Only a small proportion of our variant cohort expresses a two-fold difference from the gnomAD NFE dataset (Fig. 1c). A more pronounced difference between the datasets emerges when only considering variants annotated as (likely) pathogenic in the ClinVar database[35], where variants observed in our population deviate significantly from the gnomAD NFE populations (Fig. 1c). Additionally, the comparison of common variation exceeding 10% in the Slovenian population reveals the consistency of frequency estimates with the gnomAD NFE population, and on the other hand, considerable differences when comparing to other worldwide populations (Fig. 1b). We have additionally compared the genomic variability of the Slovenian population to other European populations that reveal its close relation to many populations from Central (Polish, Czech, Hungarian) and Western Europe (French, Utah residents with Northern and Western European ancestry - labeled as West-European, Icelandic, and Britons) (Supplementary Fig. 2).

Mitochondrial variation

As many as 3291 variants were found in the mitochondrial genome of our study cohort, where 32.1% are rare (variant frequency < 0.1%), 30.5% are uncommon (< 1%), 11.4% are common (< 10%) and 25.9% are very common (> 10%) variants. More than half of the variants are synonymous (51.3%), missense variants present 31.4% of the variants, and 16.3% are noncoding in terms of variant types. Other variants are much less frequent (Table 2).

Copy number variants

We have detected 217,553 distinct copy number variants (104,612 deletions and 112,941 duplications), cumulatively appearing over 4,4 million times in our cohort. The majority of the variants in the dataset were rare, with over half appearing only once (21,655 deletions and 58,064 duplications), and the common variants were rarely present. Still, the distribution of the different CNV (copy number variation) size groups is comparable within the frequency categories (Fig. 1d).

Pathogenic variation

We have analysed the frequency of the pathogenic variants in genes, associated with recessive disorders and are presenting the prevalence of pathogenic variants in 52 genes, exceeding 1/200 (0.5%), cumulatively covering 90.45% of the risk for recessive disorders in our population (Fig. 2). Based on the carrier frequencies detected for these genes, we have calculated the risk of having a child with a recessive disease in our population (Supplementary Fig. 3).

The highest number of pathogenic variants was detected in the *GJB2* gene (5.4%), which predicts the presence of GJB2-associated deafness in 0.29% of the population, followed by *CFTR* (5.3%) with the risk for cystic fibrosis (0.28%), *DHCR7* (3.6%) with the risk for Smith-Lemli-Opitz syndrome (0.13%), *PAH* (3.3%) with the risk for phenylketonuria (0.11%) and *CNGA1* (3.1%) with the risk for retinitis pigmentosa (0.09%). The carrier rate for *SMN1* was 1.97%, resulting in the risk for spinal muscular atrophy in 0.04% of the population. The rest of the recessive diseases have a risk of less than 0.01%. The cumulative risk of a recessive disorder in any of the genes listed below is 1.49% in our population.

Slovenian genome variant browser (SloGenVar)

We have prepared a comprehensive database of Slovenian genome variants and an open-access online variant browser, SloGenVar at https://slogenvar.si. SloGenVar browser is a user-friendly, web-based genome browser that contains data about the genomic location, frequencies, and functional consequences of genetic variants for 30,703,820 variants in 9,425 non-related individuals in the Slovenian population. The browser displays anonymized and aggregated data on the frequencies of detected variants within the Slovenian population. It is a comprehensive platform for querying and exploring genomic variants that is based on NHLBI's TOPMed program open-coded Bravo variant server[36–38]. It provides an interactive visualization of the genomic variability and detailed variant and genotype information (Fig. 3). Key features and functionalities include: variant search (users can search for variants using gene name, chromosome location, or rsID), variant details (the browser displays detailed information on each variant, including reference and alternate alleles, predicted consequence and impact, affected gene and transcript, and ClinVar annotations when available), allele frequencies (users can view allele frequencies across various populations and studies, along with a summary of the number of homozygous and heterozygous individuals for each allele), genotype quality metrics (quality metrics such as sequencing depth, genotype quality, and genotype likelihood ratio test results), data export (variant data can be downloaded in multiple formats, including VCF, JSON, and CSV), filtering options (variants can be filtered

Variant type	Mitochondrial genome Number (%)
Synonymous	1,687 (51.3)
Missense	1,032 (31.4)
Non-coding	537 (16.3)
Other	35 (1.1)

Table 2. Mitochondrial Variants.

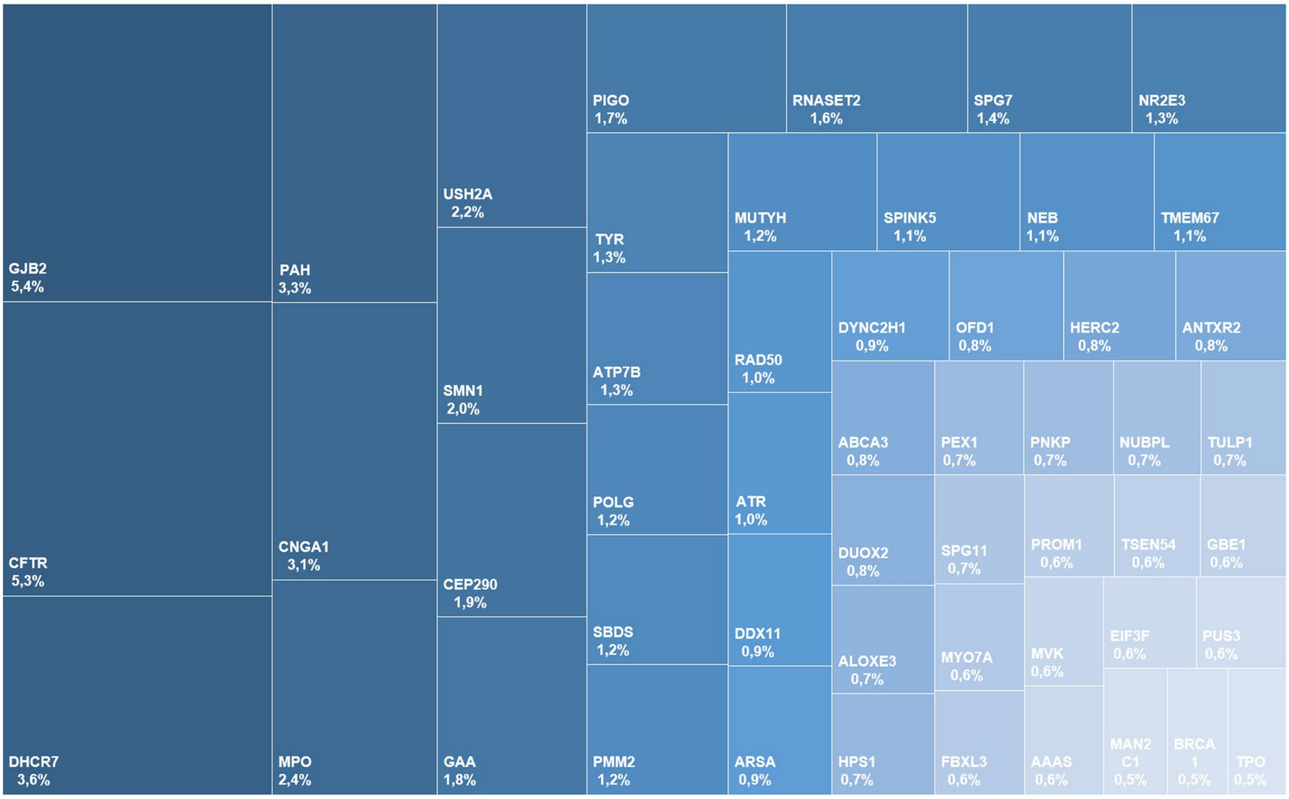

Fig. 2. Carrier frequencies in the Slovenian population. Genes with carrier rates > 1/200 are included.

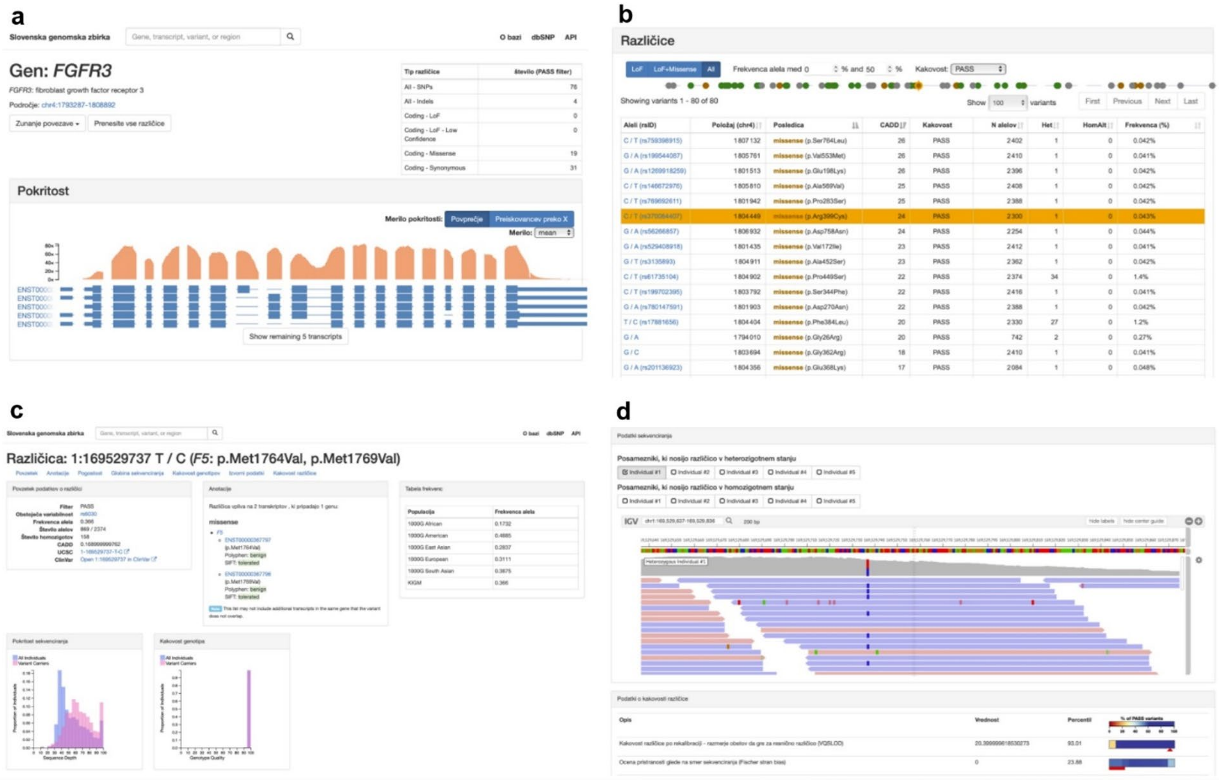
a
Slovenska genomska zbirka
O bazi
dbSNP
API
Gen: FGFR3
FGFR3: fibroblast growth factor receptor 3
Pokritost
b
Različice
c
Slovenska genomska zbirka
Različica: 1:169529737 T / C (F5: p.Met1764Val, p.Met1769Val)
d
IGV

Fig. 3. Schematic presentation of the SloGenVar browser. The panels present the following browser functionalities: (**a**) Gene-level view, including coverage information. (**b**) General variation view in a gene or a region. (**c**) Variant frequency and quality view. (**d**) Inspection of raw data for each variant/genotype is available.

by several criteria, such as allele frequency, functional consequence, impact, ClinVar annotation, and specific studies), comparison tools (the platform enables comparison of variants across populations or studies using scatter plots or tabular data), and genome browser visualization (variants can be visualized in the context of the genome, displaying the genomic region with relevant annotations).

Discussion

The present paper exhibits genomic variation in the Slovenian population, resulting in the identification of more than 30 million genomic variants, derived from sequencing 9,425 non-related Slovenian individuals, that represent more than 0.44% of the population. This work presents the biggest national genomic sequencing database in any Central European population described to date and provides a comprehensive understanding of the genomic structure of the Slovenian population.

The variability of the Slovenian population, presented in this paper, consists of single nucleotide variants, mitochondrial variants, and copy number variants. The vast majority of the detected variants were non-coding single nucleotide variants, detected in the WGS cohort, reflecting the human genome structure[39]. Most of the detected variation is extremely rare, in both SNV and CNV datasets, and over 50% of the variants appear in under 10% of the population. As the distribution of minor allele frequencies is strongly skewed towards an excess of rare variants in the human genome[2], which is an expected result, especially in a non-isolated population[26], and may be a consequence of genomic selection[40] and recent explosive growth of the human population[41]. Moreover, a significant proportion (over 3,9 million SNV variants) of the detected variation presents novel variants previously undescribed in population databases. The detection of these novel, never-before-described variants and their deposition to the public domain adds to the knowledge base of rare variation in an unaffected human population. What is more, it importantly enhances the presence of the variants, primarily detected in previously scarcely represented populations of the Central European region in publicly available control databases.

Slovenia is situated in the southern part of Central Europe, on the meeting point between the three major European language groups, and has, as such, served as a gateway for several human migrations during human history, thus indicating a historically turbulent area with many different influences and genomic contributions from the geographically neighboring as well as distant populations[42]. According to the previous research comparing the variation among the European populations, there is considerable overlap among European populations, which drops exponentially with geographic distance, and as such mirrors geographical distribution in Europe and places the Slovenian population in Central Europe[19,43,44]. This is supported by our results since most of the variation found in our population matches the frequency of the variation found in the gnomAD non-Finish European population. Additionally, we have shown a proportion of variants with more than two-fold differences in the frequency distribution and a large number of novel, never-before-described variants. The differences in distribution and the novel variation in our population may be a consequence of population-specific founder variants and other population-specific events, such as recent migrations[45]. Indeed, previous studies have shown a high number of common ancestors dating from the period of Slavic migration into Europe in the areas, that are nowadays occupied by the Slavic languages-speaking populations[19]. Moreover, a smaller study, exploring the origins of the Slovenian population has shown a close relationship of the Slovenian population to the Central European populations, and other populations, including Northern and Western European populations, are placed in a separate cluster[42]. Importantly, the gnomAD non-Finish population, used for the comparison of variant prevalence in our study, mostly consists of Western and Northern European samples[46], and hence this bias possibly explains a part of the difference between the Slovenian and the gnomAD population detected in our study. A further large-scale comparison of the Slovenian population with other European populations might reveal their relations, which may soon be feasible, as other Central European nations are beginning to prepare their own population genome sequencing projects[47].

The differences in the variant frequency distribution between the Slovenian and the gnomAD NFE populations are particularly evident when comparing ClinVar pathogenic variants. This may be explained by the possible evolutionary pressure placed on the pathogenic variants, resulting in a different distribution of pathogenic variants in different populations[48]. Indeed, previous studies have shown examples indicating that the distribution of specific pathogenic variants in the Slovenian population is different from that described in other populations[21,49].

The analysis has revealed the genes in which the pathogenic variants are most commonly present in the Slovenian population. The two genes that most frequently carry a pathogenic variant in the Slovenian population are the *GJB2* (5.4% of our population) and the *CFTR* gene (present in 5.3% of our cohort), followed by *DHCR7* (3.6%), *PAH* (3.3%), and *CNGA1* (3.1%). This distribution is different from the gnomAD NFE population, where the highest frequency of pathogenic variants among the genes associated with recessive disorders analysed in this study was discovered in the *CFTR* gene (4.6%), followed by *MPO* (2.8%), and *GJB2* and *PAH* (2.6%). Pathogenic variants in *CNGA1* were present in only 0.05% of the population[50]. Another recent study on the carrier rate of serious recessive disorders in different ethnicities by the American College of Obstetricians and Gynecologists and the American College of Medical Genetics and Genomics has explored the prevalence of pathogenic variants in the gnomAD Northern European (NE) and Southern European (SE) populations and suggests that the highest carrier rate is in the *CFTR* gene (NE: 4.6%, SE: 3.7%). GJB2 is amongst the top four in both populations (NE: 2.9%, SE: 3.2%) and SMN1 is present in a high proportion (NE: 2.1%, SE: 2.3%)[51]. These results reveal that the carrier frequency of recessive pathogenic variants in the Slovenian population differs from the gnomAD NFE and gnomAD Northern/Southern European populations.

While carrier frequencies of recessive diseases are well known to vary markedly among different ethnicities, such as large subgroups in the gnomAD project[50,51], our study reveals that population-specific differences are also measurable. For instance, *CFTR* has been previously identified as the gene with the highest frequency of carriers in the European population[50,51]. Our analysis has shown, that while pathogenic variants in *CFTR* are

Gender %		Age (years)					Disease status %	
		Min	1st quart	Median	3rd quart	Max	Affected	Unaffected
Female	51.2	0	27.4	38.4	51.8	94.7	91.3	8.7
Male	48.8	0	24.4	39.4	53.3	89.4	91.0	9.0

Table 3. Study cohort summary statistics.

highly prevalent in the Slovenian population, *GJB2* has the highest carrier frequency rate. Furthermore, the prevalence of pathogenic variants in *CNGA1* gene is present in 3.1% of our study population, which is a 30-fold increase of the estimation in the gnomAD cohort[50], suggesting that *CNGA1*-related retinitis pigmentosa is more frequent in the Slovenian population than in the gnomAD NFE population. These results may indicate that recessive gene panels selected for newborn and carrier screening should be based on population-specific carrier frequencies, rather than on the pre-defined list of conditions, that may or may not be frequent in a given population.

All the rare population variation, including pathogenic variation, detected in the Slovenian population is presented in the Slovenian genome browser SloGenVar and is freely available to the interested public upon registration. The Slovenian Genomic Browser aims to provide data for biomedical research, such as population research for discovering the genomic architecture of the Slovenian population and its relation to other European populations. Importantly, the genomic sequencing of 0.44% of Slovenians could, in the future, reveal many other Slovenian-population-specific genomic data and drive the development of precision medicine-based population health strategies, such as population-specific screening and health policy development[23].

In conclusion, we present the Slovenian genomic database and browser, the first large publicly available Central European population genomic resource available to clinical professionals from the region and beyond, to consult in the variant interpretation process. We aim to equip genomic professionals with a large control database, which includes unique Central European variants that have scarcely been included in the currently available population control databases. In light of this, the Slovenian Genome Database and browser mark a new milestone in the genomic research in the region.

Materials and methods

Study cohort, sample collection, and processing

The Slovenian genome database and browser are based on the whole genome sequencing of 619 individuals and the whole exome sequencing of 8,806 individuals of Slovenian origin. In total, 9,425 non-related individuals were carefully selected from patients recruited for genome-based diagnostics at the Clinical Institute of Genomic Medicine of the University Medical Centre Ljubljana, Slovenia between 2015 and 2024.

The inclusion criteria for our study patients were being recruited for genome-based diagnostics at the Clinical Institute of Genomic Medicine of the University Medical Centre Ljubljana, Slovenia, between 2015 and 2024, being of Slovenian ancestry, and having signed an informed consent for research inclusion. The exclusion criterion for the study was relatedness to any participant in the study cohort.

The study population consists of 51.2% females and 48.8% males, aged from 0 to 94.7 and 0 to 89.4 years, with a median age of 38.4 and 39.4 years, respectively. 91% of participants had an underlying diagnosis of a genetic condition, and 9% were unaffected (Table 3).

All the participants had signed an informed consent during their genetic consultation with a clinical geneticist and donated their blood samples. All the data has been de-reciprocally de-identified, and any personally identifiable information has been excluded.

Only non-related samples were included. Relatedness of the individuals included was assessed in two ways. Firstly, by carefully examining the familial history of participants via their medical records to identify and exclude any possible relatives of the patients, previously referred to next-generation sequencing at our institution. Secondly, the degree of relatedness of the samples included was evaluated by employing the KING software[52].

Whole genome and exome sequencing

DNA isolation, whole exome sequencing of 8,806 samples, and whole genome sequencing of 619 samples were performed at the Clinical Institute of Genomic Medicine of the University Clinical Centre Ljubljana following standardized protocols, as described previously[53–55].

After DNA isolation, library preparation for the ES analysis was performed using either Twist Core Exome v1.0, Twist Exome v2.0 or IDT xGen Exome v1.0 target captures. The GS samples were processed using the TruSeq Nano library preparation kit. All the samples were sequenced on the Illumina sequencing platform, either using the NextSeq 550, HiSeq 2000 or NovaSeq 6000 sequencers. For ES samples, the median on-target coverage amounted to 155x and for GS samples the median coverage was 34x. On average, the samples attained a high coverage over the targeted regions with an on target coverage exceeding 20x coverage for 99.9% and 95.0% targets for ES and GS samples, respectively. Median mitochondrial coverage for ES and GS samples was 354x and 3664x, respectively.

Data analysis

We employed data processing workflows and software in line with other large-scale international initiatives, including the TOPMed[36,37], 100,000 Genomes (U.K. genomic project)[56], and FinnGen (Finnish project) project[57].

The workflows for genomic data analysis were developed using the Workflow Description Language (WDL) (https://openwdl.org/), an open standard for describing data processing workflows based on the existing best practices for data analysis (GATK Best Practice guidelines)[58]. This approach ensured compliance with international standards for data analysis, facilitated the adoption of the latest workflow version, and offered scalability to the population-size data analysis of genomic data. The analytical pipelines were prepared for operation on a Slurm-based high-performance computing (HPC). All data processing has been performed on the infrastructure of Slovenian National Supercomputing Network (SLING, https://www.sling.si/en/), specifically HPC Vega, the primary supercomputer system of the Slovenian national research infrastructures upgrade project (HPC RIVR, https://www.hpc-rivr.si/home_en/). HPC Vega was delivered as the first of eight peta and pre-exa-scale EuroHPC Joint Undertaking (https://eurohpc-ju.europa.eu/index_en) systems and is hosted at the Institute of Information Science (IZUM, https://www.izum.si/en/hpc-en/) in Maribor, Slovenia. The developed workflows were versioned, and their latest versions are publicly available.

We analysed the raw sequencing data using the pipeline based on the Broad Institute's production workflow Whole Genome Germline Single Sample workflow (WDL v.3.1.11) and generated genomic VCF (multi-sample VCF) files based on the GRCh38 assembly. Multi-sample VCF files were called jointly using the Joint Genotyping workflow (WDL v.1.6.6) separately for WGS and WES cohort. Both multi-sample VCF files were then processed using bcftools v1.21[59]. Variants were left-aligned, indels were normalized, and sites were split into multiple rows. Variants other than SNP and indel were filtered out. A threshold of genotyping quality $GQ < 20$ was applied and genotypes below that threshold were marked as missing. Allele counts and frequencies were re-calculated and variants with missing genotypes in 10% or more of samples (F_MISSING > = 0.1) or zero allele counts ($AC = 0$) were excluded. Overall genotyping quality threshold of $QUAL > 100$ was applied. The thresholds for removing variants with missing genotype rates exceeding 0.1 and variant quality scores above 100 were selected as we observed that the majority of spurious variant calls due to sequencing and alignment errors were removed using these filters. Finally, multi-sample VCF files from WES and WGS cohorts were merged using the default bcftools strategy and used for the Slovenian genome database and browser, and for frequencies estimation. Finally, Ensembl Variant Effect Predictor (VEP) v.113[60] was used to annotate the variants.

To remove the bias of including the patients with genetic diagnoses in our cohort, the primary findings (pathogenic variants related to the referral diagnoses) were subsequently removed (marked as missing data) from the database and the browser.

The copy number variants (CNV) were analysed using the ExomeDepth algorithm[61] and CNV calls were considered to represent identical CNVs if their pair-wise reciprocal overlap exceeded 0.8. The mitochondrial data were analysed using the MuTect Mitochondrial workflow (https://github.com/broadinstitute/gatk/tree/master/scripts/mutect2_wdl).

To compare the genomic variability of the Slovenian population to other European populations, Euclidean distances between individual genomes were calculated and projected by Kruskal's non-metric multidimensional scaling[62]. Genomes of individuals of European ancestry were obtained from the Thousand Genomes Project[7], Human Genome Diversity Project[63] and the Allen Ancient DNA Resource (AADR) databases[64]. Genomes were limited to modern genomes and to those that share a minimum of 10% of any of the six most prominent Slovenian ancestry components in their average genome.

The Slovenian genome browser

The genomic variation, extracted from the Slovenian genome database, is publicly available at the Slovenian Genome Browser, the SloGenVar (https://slogenvar.si), an interactive, user-friendly web-based tool. Bravo DataPreparation and CoveragePreparation workflows were used to create a variant database and a browser. The implementation is based on the BRAVO variant browser from NHLBI Trans-Omics for Precision Medicine[36–38].

The SloGenVar consists of chromosome locations on GRCh38 human genome assembly, alleles, functional annotations, and allele frequencies for 30,703,820 variants called by the Joint Genotyping workflow, observed in 9,425 non-related individuals sequenced at the Clinical Institute of Genomic Medicine of the University Medical Centre Ljubljana, including 619 deeply sequenced (> 38X) genomes and 8,806 exomes. For each variant, one primary consequence was selected[60]. The browser provides an interactive visualization of the genomic variability and detailed variant and genotype information (Fig. 3).

Bravo DataPreparation and CoveragePreparation workflows were used to create a variant database and a browser. The Bravo variant browser is based on top of the MongoDB database, optimized for the storage and retrieval of large-scale genomic variant and annotation data. The core schema comprises several principal collections: variants, which contains site-specific variant data parsed from VCF/BCF files; genes, transcripts, and exons, which represent GENCODE-derived gene models; and metrics, which stores precomputed summary statistics across all variants. Additional collections such as dbsnp, users, and whitelist manage external variant identifiers and access control metadata. Variant annotations, including functional consequences and transcript-level information, are embedded within the variants collection, enabling efficient genomic interval and gene-based queries without the need for relational joins[36–38]. The minor difference between the number of SNVs detected in the Slovenian population (30,836,190) and reported in SloGenVar (30,703,820) is due to additional quality filtration applied in SloGenVar.

The SloGenVar browser enables variant search within the database by gene name, transcripts, chromosome location (variant or region), and rsID of the variant. The browser can be explored in two ways: by gathering information on a single variant or a genome region. Key information presented in the SloGenVar browser includes: variant details (reference and alternate alleles, predicted consequence and impact, affected gene and transcript, ClinVar annotations when available), allele frequencies (across various populations and studies, along with a summary of the number of homozygous and heterozygous individuals for each allele), genotype quality metrics (sequencing depth, genotype quality, genotype likelihood ratio test results), data export (in multiple formats, including VCF, JSON, and CSV), filtering options (variants can be filtered by several criteria, such as allele frequency, functional consequence, impact, ClinVar annotation, and specific studies), comparison tools (comparison of variants across populations or studies using scatter plots or tabular data), and genome browser visualization (variants can be visualized in the context of the genome, displaying the genomic region with relevant annotations).

Ethics statement

All participants provided written informed consent via an informed consent form for genetic testing as well as an informed consent form for next-generation sequencing. The research was approved by the Committee of the Republic of Slovenia for Medical Ethics (0120–286/2020/3). The study was performed in compliance with the 1964 Declaration of Helsinki and its subsequent amendments and strictly adhered to the General Data Protection Regulation (GDPR) act.

Data availability

The datasets generated and/or analysed during the current study are available in the SloGenVar browser, https://slogenvar.si. All genomic data, discovered in the Slovenian population (over 30 million of variants) have been deposited in our newly created, publicly available resource, the SloGenVar. It is free and available to anyone upon free registration with a Gmail account. The SloGenVar provides an open-access, publicly available resource to all the variants detected in our study. Since ours is an extensive collection of population data, the variants do not have accession numbers; however, all the 30 million+ variants are available for free browsing to anyone. The core message of our manuscript is the presentation of this new and important database and browser.

Received: 21 March 2025; Accepted: 16 October 2025
Published online: 20 November 2025

References

1. Goswami, C., Chattopadhyay, A. & Chuang, E. Y. Rare variants: data types and analysis strategies. *Ann. Transl Med.* **9**, 961 (2021).
2. Gibson, G. Rare and common variants: Twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2012).
3. Mathieson, I. & Reich, D. Differences in the rare variant spectrum among human populations. *PLoS Genet.* **13**, e1006581 (2017).
4. Prohaska, A. et al. Human disease variation in the light of population genomics. *Cell* **177**, 115–131 (2019).
5. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18**, 77 (2017).
6. Jakobsson, M. et al. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998–1003 (2008).
7. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
8. Bick, A. G. et al. Genomic data in the all of Us research program. *Nature* **627**, 340–346 (2024).
9. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
10. Chen, W., Coombes, B. J. & Larson, N. B. Recent advances and challenges of rare variant association analysis in the biobank sequencing era. *Front. Genet.* **13**, (2022).
11. Whiffin, N., Ware, J. S. & O'Donnell-Luria, A. Improving the Understanding of genetic variants in rare disease with Large-scale reference populations. *JAMA* **322**, 1305–1306 (2019).
12. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
13. Cavalli-Sforza, L. L. The human genome diversity project: past, present and future. *Nat. Rev. Genet.* **6**, 333–340 (2005).
14. Oleksyk, T. K., Wolfsberger, W. W., Schubelka, K., Mangul, S. & O'Brien, S. J. The pioneer advantage: Filling the blank spots on the map of genome diversity in Europe. *GigaScience* **11**, giac081 (2022).
15. Popejoy, A. B. et al. The clinical imperative for inclusivity: Race, Ethnicity, and ancestry (REA) in genomics. *Hum. Mutat.* **39**, 1713–1720 (2018).
16. Momozawa, Y. & Mizukami, K. Unique roles of rare variants in the genetics of complex diseases in humans. *J. Hum. Genet.* **66**, 11–23 (2021).
17. Lacaze, P., Manchanda, R. & Green, R. C. Prioritizing the detection of rare pathogenic variants in population screening. *Nat. Rev. Genet.* **24**, 205–206 (2023).
18. Janavičius, R. & Founder BRCA1/2 mutations in the europe: implications for hereditary breast-ovarian cancer prevention and control. *EPMA J.* **1**, 397–412 (2010).
19. Ralph, P. & Coop, G. The geography of recent genetic ancestry across Europe. *PLoS Biol.* **11**, e1001555 (2013).
20. Novembre, J. et al. Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
21. Kotnik, U., Maver, A., Peterlin, B. & Lovrecic, L. Assessment of pathogenic variation in gynecologic cancer genes in a National cohort. *Sci. Rep.* **13**, 5307 (2023).
22. Abul-Husn, N. S. et al. Implementing genomic screening in diverse populations. *Genome Med.* **13**, 17 (2021).
23. Molster, C. M. et al. The evolution of public health genomics: exploring its past, present, and future. *Front. Public. Health* **6**, (2018).
24. Kovanda, A., Zimani, A. N. & Peterlin, B. How to design a National genomic project-a systematic review of active projects. *Hum. Genomics*. **15**, 20 (2021).
25. Gurdasani, D. et al. Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell* **179**, 984–1002e36 (2019).
26. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
27. Naslavsky, M. S. et al. Whole-genome sequencing of 1,171 elderly admixed individuals from São Paulo, Brazil. *Nat. Commun.* **13**, 1004 (2022).
28. Jain, A. et al. IndiGenomes: a comprehensive resource of genetic variants from over 1000 Indian genomes. *Nucleic Acids Res.* gkaa923 https://doi.org/10.1093/nar/gkaa923 (2020).
29. Kamada, M. et al. MGeND: an integrated database for Japanese clinical and genomic information. *Hum. Genome Var.* **6**, 53 (2019).
30. Li, Z. et al. CMDB: the comprehensive population genome variation database of China. *Nucleic Acids Res.* **51**, D890–D895 (2023).
31. Reis, A. L. M. et al. The landscape of genomic structural variation in Indigenous Australians. *Nature* **624**, 602–610 (2023).
32. NCMG. database of genomic variants.
33. Kaja, E. et al. The thousand Polish Genomes-A database of Polish variant allele frequencies. *Int. J. Mol. Sci.* **23**, 4532 (2022).
34. Slovenski genomski projekt. https://cris.cobiss.net/ecris/si/sl/project/17959

35. Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, 980–985 (2014).
36. The, N. H. L. B. I. *Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Program* (BRAVO variant browser, 2018).
37. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI topmed program. *Nature* **590**, 290–299 (2021).
38. Burgess, D. J. The topmed genomic resource for human health. *Nat. Rev. Genet.* **22**, 200–200 (2021).
39. Boland, R. & Non-coding, C. It's not junk. *Dig. Dis. Sci.* **62**, 1107–1109 (2017).
40. Li, Y. et al. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.* **42**, 969–972 (2010).
41. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740–743 (2012).
42. Maisano Delser, P., Ravnik-Glavač, M., Gasparini, P., Glavač, D. & Mezzavilla, M. Genetic landscape of slovenians: past admixture and natural selection pattern. *Front. Genet.* **9**, 1–8 (2018).
43. Busby, G. B. J. et al. The role of recent admixture in forming the contemporary West Eurasian genomic landscape. *Curr. Biol.* **25**, 2518–2526 (2015).
44. Nelis, M. et al. Genetic structure of europeans: a view from the North-East. *PLoS One*. **4**, e5472 (2009).
45. Jain, A., Sharma, D., Bajaj, A., Gupta, V. & Scaria, V. Chapter Four - Founder variants and population genomes—Toward precision medicine. In *Advances in Genetics* (ed Kumar, D.) vol 107 121–152 (Academic, (2021).
46. Karczewski, K. J. et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv* 531210 https://doi.org/10.1101/531210 (2019).
47. Smetana, J. & Brož, P. National genome initiatives in Europe and the united Kingdom in the era of Whole-Genome sequencing: A comprehensive review. *Genes (Basel)* 13, (2022).
48. Dudley, J. T. et al. Human genomic disease variants: a neutral evolutionary explanation. *Genome Res.* **22**, 1383–1394 (2012).
49. Vodnjov, N. et al. A novel splice-site FHOD3 founder variant is a common cause of hypertrophic cardiomyopathy in the population of the Balkans-A cohort study. *PLoS One*. **18**, e0294969 (2023).
50. Zhu, W. et al. A robust pipeline for ranking carrier frequencies of autosomal recessive and X-linked Mendelian disorders. *Npj Genomic Med.* **7**, 72 (2022).
51. Johansen Taber, K. et al. A guidelines-consistent carrier screening panel that supports equity across diverse populations. *Genet. Med.* **24**, 201–213 (2022).
52. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
53. Bergant, G. et al. Comprehensive use of extended exome analysis improves diagnostic yield in rare disease: A retrospective survey in 1,059 cases. *Genet. Med.* **20**, 303–312 (2018).
54. Kovanda, A. et al. A multicenter study of genetic testing for parkinson's disease in the clinical setting. *NPJ Parkinsons Dis.* **8**, 149 (2022).
55. Bergant, G., Maver, A. & Peterlin, B. Whole-Genome sequencing in diagnostics of selected Slovenian undiagnosed patients with rare disorders. *Life (Basel)* **11**, (2021).
56. The 100,000 Genomes Project Pilot Investigators. 100,000 genomes pilot on Rare-Disease diagnosis in health Care — Preliminary report. *N Engl. J. Med.* **385**, 1868–1880 (2021).
57. Kurki, M. I. et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
58. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinf.* **43**, 11101–111033 (2013).
59. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* 10, (2021).
60. Ensembl Ensembl Variation - Calculated variant consequences. https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html (2024).
61. Ellingford, J. M. et al. Validation of copy number variation analysis for next-generation sequencing diagnostics. *Eur. J. Hum. Genet.* **25**, 719–724 (2017).
62. Hout, M. C., Papesh, M. H. & Goldinger, S. D. Multidimensional scaling. *Wiley Interdiscip Rev. Cogn. Sci.* **4**, 93–103 (2013).
63. Tontonoz, M. The Human Genome Diversity Project (1991–2002). *Embryo Project Encyclopedia* (2025).
64. Mallick, S. et al. The Allen ancient DNA resource (AADR) a curated compendium of ancient human genomes. *Sci. Data*. **11**, 182 (2024).

Acknowledgements

We acknowledge the members of the Slovenian National Supercomputing Network (SLING, https://www.sling.si/en/https://www.sling.si/en/) for their support and collaboration during the population-scale re-analysis of genomic data. We are especially grateful to Barbara Krašovec for her invaluable assistance with the HPC infrastructure.

Author contributions

AM, PJ, and UK contributed equally to the manuscript and share the first authorship. AM has established the Slovenian Genome Project, analysed the variation in the Slovenian population and prepared the graphical results. PJ analysed the primary data and established the SloGenVar database and the variant browser. UK has written the first draft of the manuscript. BP, AM, and PJ conceived and designed the study. UK, AM, PJ, and BP contributed to the interpretation of the results. LL and GB contributed to the design of the research. All authors read and approved the final manuscript.

Funding

This work was funded by the ARRS project V3-1911: Slovenian genome project.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-24991-9.

Correspondence and requests for materials should be addressed to B.P.

Reprints and permissions information is available at www.nature.com/reprints.