‘The Thousand Polish Genomes Project’ - a national database of Polish variant allele frequencies

*Elżbieta Kaja^{1,5}, *Adrian Lejman^{1,5}, Dawid Sielski¹, Mateusz Sypniewski^{1,10}, Tomasz Gambin³, Tomasz Suchocki^{2,13}, Mateusz Dawidziuk⁹, Paweł Golik⁴, Marzena Wojtaszewska^{1,5}, Maria Stępień^{1,11}, Joanna Szyda^{2,13}, Karolina Lisiak-Teodorczyk¹, Filip Wolbach¹, Daria Kołodziejska¹, Katarzyna Ferdyn^{1,12}, Alicja Woźna^{1,8}, Marcin Żytkiewicz⁷, Anna Bodora-Troińska⁷, Waldemar Elikowski⁷, Zbigniew Król⁵, Artur Zaczynski⁵, Agnieszka Pawlak^{5,14}, Robert Gil⁵, Waldemar Wierzbą⁵, Paula Dobosz^{1,6}, Katarzyna Zawadzka¹, Paweł Zawadzki^{1,8}, Paweł Sztromwasser¹

¹ MNM Diagnostics Sp. z o.o., ul. Macieja Rataja 64, Poznań, 61-695, Poland

² Biostatistics Group, Wrocław University of Environmental and Life Sciences, Wrocław, Poland

³ Institute of Computer Science, Warsaw University of Technology, Warsaw, 00-665, Poland

⁴ Institute of Genetics and Biotechnology, Faculty of Biology, University of Warsaw, Warsaw, 02-106, Poland

⁵ Central Clinical Hospital of Ministry of the Interior and Administration in Warsaw, Warsaw, 02-507, Poland

⁶ Department of Hematology, Transplantation and Internal Medicine University Clinical Center of the Medical University of Warsaw, 02-091, Warsaw, Poland

⁷ Internal Diseases Department, Józef Struś Multidisciplinary Municipal Hospital, Poznań, 61-285, Poland

⁸ Molecular Biophysics Division, Faculty of Physics, A. Mickiewicz University, Poznań, 61-614 Poland

⁹ Department of Medical Genetics, Institute of Mother and Child, Warsaw, 01-211, Poland

¹⁰ Department of Genetics and Animal Breeding, Poznań University of Life Sciences, Wołyńska 33, 60-637 Poznań, Poland

¹¹ Faculty of Medicine, Medical University of Lublin, 20-059 Lublin, Poland

¹² Medical and Science Sp. z o.o., Pódeblacie 107B, 08-455, Poland

¹³ National Research Institute of Animal Production, Krakowska 1, 32-083 Balice, Poland

¹⁴ Mossakowski Medical Research Centre, Polish Academy of Science, Pawińskiego 5, 02-106 Warsaw, Poland

*of equal contribution

Corresponding author:

Paweł Sztromwasser

pawel.sztromwasser@mmn.bio

Keywords: gene variant, allele frequency, genome, whole genome sequencing, population genomics, allelic distribution, Polish genomes, national database

Abstract

Although Slavic populations account for over 3.5% of world inhabitants, no centralized, open source reference database of genetic variation of any Slavic population exists to date. Such data are crucial for either biomedical research and genetic counseling and are essential for archeological and historical studies. Polish population, homogenous and sedentary in its nature but influenced by many migrations of the past, is unique and could serve as a good genetic reference for middle European Slavic nations.

The aim of the present study was to describe first results of analyses of a newly created national database of Polish genomic variant allele frequencies. Never before has any study on the whole genomes of Polish population been conducted on such a large number of individuals (1,079).

A wide spectrum of genomic variation was identified and genotyped, such as small and structural variants, runs of homozygosity, mitochondrial haplogroups and Mendelian inconsistencies. The allele frequencies were calculated for 943 unrelated individuals and released publicly as The Thousand Polish Genomes database. A precise detection and characterisation of rare variants enriched in the Polish population allowed to confirm the allele frequencies for known pathogenic variants in diseases, such as Smith-Lemli-Opitz syndrome (SLOS) or Nijmegen breakage syndrome (NBS). Additionally, the analysis of OMIM AR genes led to the identification of 22 genes with significantly different cumulative allele frequencies in the Polish (POL) vs European NFE population. We hope that The Thousand Polish Genomes database will contribute to the worldwide genomic data resources for researchers and clinicians.

Introduction

An individual genome carries about four million single nucleotide variants and indels, and thousands of structural variants. These alterations in the DNA are mostly responsible for 0.1% of difference between genomes of two unrelated individuals (The 1000 Genomes Project Consortium 2015). Genetic variation is the reason for phenotypic differences - disease susceptibility, drug responses and physical traits such as height (Frazer et al. 2009). Mapping naturally occurring genetic variation is thus critical for studying human health and disease.

The completion of the Human Genome Project and subsequent large-scale projects, such as the HapMap (†The International HapMap Consortium 2003), the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015), The ExAC Project (Exome Aggregation Consortium et al. 2016) and the TopMED (NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium et al. 2021) provided large datasets for human genomic variation studies. Most of these efforts were focused on the largest populations (<https://sci-hub.se/https://www.nature.com/articles/nature18964>) and dominated by participants of caucasian ancestries, so despite sheer sizes, the information on the full spectrum of world's genetic variation remains incomplete. With short read sequencing technology becoming cheaper and more accessible across the world, new large scale initiatives are launched, targeting thousands of genomes to fill in the gaps in our knowledge and build stronger foundations for precision medicine. In Europe, 'The 1+ Million Genomes initiative' (<https://b1mg-project.eu/>) is aiming to create the most comprehensive map of European diversity by bolstering up separate efforts of EU countries to sequence their populations. Similar projects are being held in the USA, 'All of us' project (<https://allofus.nih.gov/>) in Asia, GenomeAsia 100K Project (Wall et al. 2019) and Africa, 3MAG Project (ambitious Three Million African Genomes). Genetic variants derived from 'healthy' genomes on population scale are essential as a reference data for studying diseases, epidemiology, human diversity, evolution and migration. For example, the ClinVar (<http://www.ncbi.nlm.nih.gov/clinvar/>), Varsome (<http://varsome.com>) and HGMD (<http://www.hgmd.org>) databases store information about

pathogenicity status of millions of small human variants (single nucleotide variants (SNVs) and indels), categorizing them to 5 groups according to ACMG guidelines (benign, likely benign, VUS, likely pathogenic and pathogenic). This clinical classification could not be created without population scale variant databases such as the 1000 Genomes or gnomAD (<https://gnomad.broadinstitute.org>). Similarly, Database of Genomic Variants (DGV) and DECIPHER aggregate the Structural Variants (SVs) and Copy Number Variants (CNVs) information for public use.

Here, we report the insights from the Thousand Polish Genomes (POL) project, an effort to sequence and characterize 1079 whole genomes of Polish individuals, including over 110 families. High quality and depth of the sequencing data (the average coverage above 35x) provide a unique repository of genetic variation in the Polish population. We characterized a wide spectrum of genetic alterations, including small and structural variants, as well as mitochondrial DNA. Taking into account that populations of central-European ancestry are underrepresented in the world's large-scale genomic databases, the presented work complements international efforts to map human genetic variation. Additionally, an open allele frequency database¹ will become an invaluable resource for clinical and population genetics, and biomedical research and demographic inference.

Materials and methods

Donors characteristics and sample collection

The studied population consisted of participants recruited within the “Search for genomic markers predicting the severity of the response to COVID- 19” project related to genetic predisposition to COVID-19 severity. Samples were collected from 1079 individuals of Polish origin between April 2020 and April 2021. The whole cohort comprises three groups: (1) 441 participants from the Central Clinical Hospital of the Ministry of Interior and Administration in Warsaw. Majority of them (391 out of 441) were patients suffering from COVID-19 caused by SARS-CoV-2 coronavirus, peripheral blood samples from this group were collected during hospitalization. (2) 608 individuals were self-enrolled volunteers for the project. Samples from this group were collected in blood collecting facilities from all over Poland. Almost all Polish local administrative units (15 out of 16 voivodeships) were represented (Figure 1). (3) 30 individuals were collected in the Multidisciplinary Municipal Hospital of Józef Strus in Poznań from patients suffering from COVID-19. The study participants were unrelated individuals (N=943) or families. Among participating families (N= 117) there were 69 trios, 12 quartets, 5 pairs of siblings and 31 families consisting of one parent and a child.

From all participants of “The Thousand Polish Genomes Project” (POL) basic clinical history was collected in order to exclude genetically affected individuals. Data collected for all individuals included: gender, age, BMI, comorbidities (diabetes, hypertension, ischemic heart disease, stroke, heart failure, cancer, kidney failure/disease, hepatitis B, chronic obstructive pulmonary disease).

Additional clinical data collected for 608 participants, not hospitalised for COVID-19 included: genetic disorders, flu vaccine, tuberculosis and measles vaccine, smoking habits, hepatitis C infection and others.

Peripheral blood samples from participants were collected in the EDTA-containing tubes. Whole process of the collection of biological material was monitored and performed in accordance with the protocol of blood collection, which is part of the Quality Management System.

¹ <https://github.com/MNMdiagnostics/1000PolishGenomes>

Compliance with Ethical Standards

All participants, or their guardians/parents (when the participants were under 18), provided their informed consent before collecting their blood samples and completed a clinical data form, which included a questionnaire about country of origin and chronic diseases. Only individuals without diagnosed severe genetic disorders were qualified for further studies. All procedures performed in studies involving human participants were in accordance with highest ethical standards, the Ethical approval for the study was obtained from Ethics Committee of the Central Clinical Hospital of the Ministry of Interior and Administration in Warsaw Ethics Committee (decision nr: 41/2020 from 03.04.2020 and 125/2020 from 1.07.2020). The study was in accordance with the 1964 Helsinki declaration and its later amendments and adhered to the highest data security standards of ISO 27001. The study is also compliant with the General Data Protection Regulation (GDPR) act.

Total Quality Management utilized in the study

The project was carried out in accordance with the Total Quality Management (TQM) methodology, which directly translates into the quality of results.

Before starting the project, the details were mapped - critical points were defined, such as reference ranges for collected biological material, its preparation, isolation, DNA concentration and quality, genomic sequencing (including quality control - QC). The legal and ethical transparency of the entire project was ensured, including confidentiality and data integrity as well as impartiality. Additionally, risk analysis was performed, possible difficulties were mapped and precise corrective actions have been planned.

Figure 1. The map of 637 individuals enrolled in “The Thousand Polish Genomes Project” (related and unrelated ones) in respect of Polish voivodeships. It refers to individuals who volunteered to the project, from whom samples were collected in blood collection facilities from all over Poland.

Methods

Whole genome sequencing

High depth (>30x) PCR-free whole-genome sequencing was applied to all samples. Genomic DNA was isolated from the peripheral blood leukocytes using a QIAamp DNA Blood Mini Kit, DNeasy Blood & Tissue Kit (Qiagen), Blood/Cell DNA Mini Kit (Syngen) and Xpure Blood Kit (A&A Biotechnology) according to the manufacturer's protocols. The concentration and purity of isolated DNA was measured using the NanoDropTM spectrophotometer, the DNA integrity was evaluated with electrophoresis. The sequencing library was prepared by Macrogen Inc. (Seul, South Korea) using TruSeq DNA PCR-free kit (Illumina Inc., San Diego, California, United States) and 550 bp inserts. Quality of DNA libraries were measured using 2100 Bioanalyzer from Agilent Technologies. Subsequently Whole Genome Sequencing (WGS) was performed on the Illumina NovaSeq 6000 platform using 150 bp paired-end reads, with median average depth of 35.72X (Supplementary Table S1).

WGS data processing

Quality of the sequenced reads was confirmed using FastQC v0.11.7² and the reads were subsequently mapped to the GRCh38 human reference genome using Speedseq framework v.0.1.2 (Chiang et al. 2015), encompassing BWA MEM 0.7.10 (H. Li 2013) mapping, SAMBLASTER v0.1.22 (Faust and Hall 2014) duplicate removal, and Sambamba v0.5.9 (Tarasov et al. 2015) sorting and indexing. Mapping coverage was calculated using Mosdepth 0.2.4 (Pedersen and Quinlan 2017) and included reads with MQ>0. Single nucleotide variants and short indels in the nuclear genome were detected using DeepVariant 0.8.0 (Poplin et al. 2018) and jointly genotyped with GLnexus (Yun et al. 2021). Next, multiallelic variants were decomposed and normalized using BCFtools (Danecek et al. 2021). Small variants in the mitochondrial genome were called using Freebayes v.0.9.21 (Garrison and Marth 2012), and mitochondrial haplogroups inferred using Haplogrep2 software (Weissensteiner et al. 2016). Structural variants were called and jointly genotyped with Smoove v.0.2.6³.

All variants were annotated using Ensembl Variant Effect Predictor v.97 (McLaren et al. 2016), including references to databases of genomic variants from ClinVar v. 201904 (Landrum et al. 2018) and dbSNP build 151, variant population frequencies from the 1000 Genomes Project (Auton et al. 2015), GnomAD v2.0.1 and v3.0 (Karczewski et al. 2020), and pathogenicity scores including Polyphen-2 (Adzhubei et al. 2010), SIFT (Sim et al. 2012), and DANN (Quang, Chen, and Xie 2015). All WGS data processing was automated with a Ruffus pipeline framework (Goodstadt 2010), and some analysis steps were parallelized with GNU parallel tool (Tange 2011).

Variant analysis

Small variants of 943 unrelated individuals were extracted from the normalized multisample VCF file using BCFtools (Danecek et al. 2021). Sites with genotyping rate below 90% and invariant in the group were excluded. The resulting dataset, with average call rate per individual of 99%, was used in all subsequent analyses. Joint called structural variants were also filtered for related individuals, and

² <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

³ <https://github.com/brentp/smoove>

subsequently annotated and filtered according to Smoove⁴ authors' recommendations, i.e. heterozygous calls with MSHQ ≥ 3 , deletions with DHFFC ≥ 0.7 , and duplication calls with DHFFC ≤ 1.25 were excluded from the analysis. We also required at least one split-read per allele in the cohort. Variant statistics were generated with BCFtools (Danecek et al. 2021), and custom scripts. Data processing and visualization was programmed in R (R Core Team 2020) with the tidyverse package (Wickham et al. 2019). Small and structural variant sets along with allele frequencies were released as a publicly available resource - The Thousand Polish Genomes database⁵.

Mendelian inconsistencies

Small variants violating Mendelian inheritance pattern were identified in 93 child-parents trios, and subsequently filtered using BCFtools (Danecek et al. 2021). For *de novo* candidates we selected sites covered by at least 10 reads in all family members, with minimum 5 alternative allele reads in the child and 0 such reads in the parents, and with minimum alt-allele depth ratio of 0.25.

Runs of homozygosity

Runs of homozygosity were identified using a hidden Markov model approach (Narasimhan et al. 2016) implemented in BCFtools (Danecek et al. 2021). We analysed RoH Identified on autosomal chromosomes, requiring more than 50 markers per RoH, and average marker quality above 25.

Recessive carriers identification

The list of Autosomal Recessive disease-associated genes was obtained from the OMIM database⁶. Variants located within these genes were extracted from the POL and gnomAD v3.0 cohorts and annotated using VEP (release 101), including ClinVar annotation (release date 17-05-2021). Of those, we selected rare (gnomAD_AF < 0.01%), nonsynonymous (VEP IMPACT = 'HIGH' or 'MODERATE'), pathogenic or likely pathogenic (according to ClinVar) variants. For each gene, we calculated cumulative allele frequencies of previously selected variants in POL and gnomAD NFE cohorts. To test whether the cumulative frequencies in particular genes differ between both cohorts we used chi-square test and applied Bonferroni correction for multiple testing to get adjusted P-values.

We downloaded gene panels data from panelapp website⁷ and selected high confidence gene-disease associations (annotated as "Expert Review Green") with ModeOfInheritance defined as "BIALLELIC". For each panel we calculated the sum of cumulative allele frequencies of pathogenic/likely pathogenic variants (obtained in the previous step) in genes included in the panel. Similarly to the gene level analysis we tested the difference between POL and gnomAD NFE populations using a chi-square test followed by Bonferroni correction.

⁴ <https://github.com/brentp/smoove>

⁵ <https://github.com/MNMdiagnostics/1000PolishGenomes>

⁶ <https://www.omim.org/>, accession date 26-04-2021

⁷ <https://panelapp.genomicsengland.co.uk/panels/>, accession date 27-06-2021

Population analysis

We evaluated the homogeneity of the Polish population (POL) using three different approaches. In the first step we checked whether all the analyzed POL samples belong to the European population. We used common variants occurring in the 1000G project and our cohort data and further pruned them to Minor Allele Frequency (MAF) over 1% and to be in linkage equilibrium. In the second step, we performed Principal Components Analysis (PCA) on a 1000G dataset and projected them into POL genotype data. We trained a random forest model with the first six principal components (PCs) on the 1000G dataset and predicted the ancestry of POL individuals. In the third step, in order to find out which European population is the closest to POL we used the subpopulations information available in 1000G data set (Utah Residents (CEPH) with Northern and Western European ancestry [CEU], Finnish in Finland [FIN], British in England and Scotland [GBR], Iberian population in Spain [IBS], Toscani in Italy [TSI]) and repeated the PCA and prediction based on random forest analyses as was described above. The only exception was to use the first two PCs in training and prediction of the random forest.

We also performed population structure analysis using ADMIXTURE (Alexander, Novembre, and Lange 2009) with the unrelated individuals from our cohort, filtered and pruned as described above, with the populations from the 1000G dataset, as well as with the five European populations. The analysis was performed in the unsupervised mode for K ranging from 2 to 7, and cross-validation error was used to determine the best value for K.

To calculate the average fixation index between two groups of individuals from SNPs we used Fst statistics calculated using the formula proposed by Weir and Cockerham (Weir and Cockerham 1984). The calculations were performed in Plink software (Purcell et al. 2007). We also compared the Polish population with all populations available within the 1000 Genomes Project and checked whether Fst patterns were consistent with the continental and the populational clustering and with previously obtained estimates. We also calculated an average Fst statistics over each sliding window built from 1,000 SNPs. We used the two first PCs and the Density-Based Spatial Method implemented in Python to determine potential subpopulations within the Polish population.

Results

Characteristics of the POL cohort

Median age of participants was 45 (2-98), with slight predominance of males (618 vs 461). The analysis of clinical data showed that the most common chronic diseases reported by participants of the POL cohort were: hypertension (13%), cancer (4,6%), diabetes (4%) and hypothyroidism or Hashimoto's disease (3%). No health problems (excluding COVID-19 infection which was a major focus of patients recruitment) were reported by 86% of participants. The whole cohort consists of **1079** individuals of Polish origin, **943** of which are unrelated. The figure below (Fig. 2) shows the age distribution of all unrelated participants.

Figure 2: Age distribution of unrelated study participants.

Genetic variation in the Polish population

We processed sequencing data from 1079 individuals, totalling 878.4 billion read-pairs corresponding to an average 35.72x mean coverage per genome (Supplementary Fig. S1; Supplementary Table S1). In all samples 92% of the genome was covered with at least 10 reads, and in 41% of samples 90% of the genome was covered by at least 20 reads. We detected and genotyped small and structural variants, runs of homozygosity, mitochondrial haplogroups, and Mendelian inconsistencies, all of which are described in detail below. Except for Mendelian inconsistencies, in all analyses we included data for 943 unrelated individuals, a group referred to as the POL cohort below. Allele frequencies of small and structural variants from this cohort constitute The Thousand Polish Genomes database⁸, that we release openly for academic research.

Small and structural variants

We identified a group of small variants including a total of 31.24M single nucleotide variants (SNV) and 5.63M small insertions and deletions. An individual genome contained on average 3.71M (3.63 - 3.77) single nucleotide substitutions of which 2.22M (2.08 - 2.36) were heterozygous and 1.49M (1.38 - 1.55) homozygous. Together with, on average, 0.77M (0.75 - 0.78) short insertions and deletions, we found 4.48M small variants per individual, of which 16,473 were private variants (Supplementary Table S2).

An average individual genome also carried 2,602 large deletions (2,431-2,724), 106 duplications (86-128), and 76 inversions (56-91). The Thousand Polish Genomes database holds a total

⁸ <https://github.com/MNMdiagnostics/1000PolishGenomes>

of 59,823 structural variants, of which 21,492 were marked as high quality (18,971 deletions, 2,148 duplications, and 373 inversions; Supplementary Table S3 and S4). Majority of deletions were shorter than 1kb (63.6%; 93.1% were <10kb), with peaks at 50bp and 300bp (Supplementary Fig. S2 A-C). Distribution of duplication sizes showed two peaks, at 200bp and at 10kb. Only 10 duplications (0.5%) exceeded 1Mb in length. Inversions, being the most rare category of the three structural variant types, displayed a clear length peak below 1kb, and the highest fraction of events longer than 1Mb (2.7%; N=10). Median length of the genomic sequence affected by SVs in a single individual was 12.6Mb (2.94Mb, 2.34Mb, and 7.34Mb for deletions, duplications, and inversions respectively), which exceeds the total DNA content affected by SNVs.

Distribution of allele-frequencies

We compared distribution of allele frequencies among different variant types: substitutions, short indels, and structural variants. The results presented in Fig. 3 indicate that the majority of substitutions (50.8%) were private variants, observed in a single individual (MAF ~0.05%). For deletions and duplications singletons constituted 41 and 43.7% respectively, while for the inversions only 26%. Common (MAF>1%) structural variants in our cohort were most abundant among inversions (50%), and least abundant among duplications (15.6%). High frequency small variants fell in the middle with 29.4% common for SNVs and 38.7% for indels. Interestingly, fixed alleles (allelic frequency equal to 100%) were present among all three types of structural variants (0.1-0.5% of all SVs) and completely absent among small variants. A summary of small variant counts with breakdown into three allele frequency tiers is shown in Table 1.

Figure 3: Distribution of allele frequencies across different variant types.

VARIANT_CLASS	AF	IMPACT			
		HIGH	LOW	MODERATE	MODIFIER
deletion	>0.5%	500	1 090	680	1 697 005
insertion		327	1 280	630	1 934 758
SNV		1 446	45 903	39 096	15 294 308
deletion	0.1-0.5%	813	582	859	885 811
insertion		407	648	573	1 010 671
SNV		1 723	32 131	41 239	9 509 163
deletion	<0.1%	2 600	946	1 684	1 383 954
insertion		1 219	773	983	1 062 161
SNV		4 967	68 293	100 838	19 225 909

Table 1: Summary of variant counts in three population allele frequencies.

Denovo variation in family-based analysis

Using small variant genotypes in 93 parents-child trios we identified on average 11,348 (10,128-13,224) Mendelian inconsistencies per trio. Considering only variants with MAF<1% in the entire cohort, a child carried on average 347 substitutions (284-456), and 1,127 indels (977-1,376). A total of 194 rare *denovo* SNVs affected the protein coding sequence (114 missense, 53 synonymous, 23 splice-site or region, and 4 stop-gain) corresponding to 0-4 exonic SNVs in 89 trios, and more than 4 in four trios (5-8). On average, a child carried 2,1 *denovo* SNVs with MAF<1%.

Runs of homozygosity

On average, an individual genome contains 1.1 runs of homozygosity (RoH) longer than 100kb, spanning a total of 284Mb. RoH were uniformly distributed along the chromosomes, with the exception of centromeric and near-centromeric regions (Fig. 5). Mean size of a RoH was 252.7kb (100kb-14Mb), and counts and average sizes of RoH on particular chromosomes can be found in Supplementary Fig. S3 and Supplementary Fig. S4. The steepest growth in cumulative length of RoHs (Fig. 4) was observed between 20kb and 200kb, and RoH longer than 200kb represented less than a quarter of the homozygous sequence (median 88Mb).

Figure 4: Cumulative length of RoHs.

Figure 5: Chromosomes coverage by runs of homozygosity in the POL cohort.

Mitochondrial haplogroups

Using variant calls in the mitochondrial genome, we inferred haplogroups among the unrelated individuals. In 930 individuals with high quality haplogroup assignment the most abundant haplogroup was H with 410 (44.1%) representatives, U with 161 (17.3%), J with 92 (9.9%), and T with 83 (8.9%) individuals (Fig. 6). The largest H sub-haplogroup was H1 (N=128; 31.2% of the H haplogroup), and a similar number of individuals was divided between subclades H2, H5, H6 and H11 (N=116; together 28.3% of the H haplogroup). The second most abundant sub-haplogroup in the cohort was U5 with 98 (10.5%) individuals.

Figure 6: Haplogroups distribution in POL.

Functional impact of the variants

Impact of small variants on protein function

Depending on the impact on the genomic sequence, variants were distributed non uniformly across the allele-frequency spectrum. Variants in protein coding regions were depleted in higher frequencies (Supplementary Fig. S5) compared to non-coding variation (e.g. intergenic, intronic). Larger differences were observed among exonic variants (Fig. 7), where protein altering variation (nonsense variants, frameshift indels, and missense variants) was associated with lower allele frequencies in contrast to UTR, synonymous, and non-exonic variants. A summary of variant counts, SNVs and indels, divided into three population frequency tiers and categorized by functional impact is presented in Table 1.

Figure 7. Distribution of exonic variant consequences across allele frequency spectrum.

Impact of rare variants on diseases

In order to infer the burden of disease causing variants in the Polish population, we considered three rare variant sets: (1) variants annotated in the ClinVar database as pathogenic or likely pathogenic; (2) high-impact variants in genes from the ACMG gene list; (3) variants associated with autosomal recessive (AR) diseases according to the OMIM database⁹. The details of the three sets and the lists of corresponding genes are described below.

In the first approach, we identified 736 rare variants (MAF<0.1% in gnomAD v3.1) reported in ClinVar as pathogenic or likely pathogenic. Selected variants were divided into four groups by the review status, which is the assertion of clinical significance: (1) 282 variants with a low confidence level (provided by a single submitter, interpreted with assertion criteria and evidence or provided by multiple submitters, but with conflicting interpretations of pathogenicity); (2) 432 variants with higher clinical significance (submitted to ClinVar by two or more submitters with the same interpretation), for example rs11555217 in *DHCR7* gene, described in more details below; (3) 20 pathogenic variants with high clinical significance (variants reviewed by an expert panel) and (4) 2 variants with the highest level of significance (assessed as “practice guideline” assertions). 533 (72.4%) of these variants were observed in a single person and 203 variants were carried by 2-22 individuals (for instance variants in *PAH*, *OTOG* and *CBS* gene).

In the second approach, we evaluated clinically relevant ‘incidental findings’ variants according to novel American College of Medical Genetics and Genomics (ACMG) recommendations. The most recent version of ACMG (v.3.0) guidelines was published in May 2021 (ACMG Secondary Findings Working Group et al. 2021). Using this list we identified 50 variants in POL having a high impact on genes. After manual curation by certified medical genetic scientists, 13 variants were assessed as pathogenic and 3 as likely pathogenic, affecting 18 of 943 (1.9%) unrelated healthy individuals in our cohort. 15 (94%) of these variants were observed in a single person and were mostly related to cancer predisposition genes such as *MUTYH*, *MSH6*, *BRCA2*, and *PALB2*. In one case the variant was carried

⁹ <https://www.omim.org/>, accession date 26-04-2021

by three individuals (rs1217805587 in *BRCA1* gene). Frequencies of these variants did not differ significantly from the gnomAD NFE population.

In the third approach, we extracted 2 512 genes associated with Autosomal Recessive (AR) diseases and then gene panels containing AR disease genes from the OMIM database. We found 26 288 and 73 036 rare (AF < 0.01 % in GnomAD), nonsynonymous (IMPACT= 'HIGH' or 'MODERATE') variants in the POL and GnomAD cohorts respectively. Of those, 705 (POL) and 10 116 (gnomAD) variants have been annotated in ClinVar database as pathogenic or likely pathogenic. Per-gene analysis of the cumulative allele frequencies of these variants revealed the existence of significant differences (Chi-square test, P value < 0.01) in 22 genes between POL and gnomAD NFE populations (Supplementary Table S5, Figure 8).

A gene with one of the the highest log2 fold change of over 3.3 in the POL cohort was *XPA*, encoding a protein that plays a crucial role in the nucleotide excision repair process, a specialized type of DNA repair (Musich, Li, and Zou 2017). Pathogenic variants in this gene cause Xeroderma Pigmentosum (XP) type A, one of seven complementation groups of this disorder. There is a variability in prevalence within complementation groups with *XPC* being the most common complement type in the United States, Europe, and North Africa while *XPA* is the most common type in Japan (Black 2016). Our results suggest that *XPA* pathogenic variants are more common in the Polish population vs gnomAD NFE populations.

Other genes with pathogenic variants that were overrepresented in the Polish populations include: *C2*, *TGM5*, *C8B*, *CBS*, *OTOG*, *GP6*, *MASP1*, *C19orf12*, *SORD*, *ALOX12B*, *FKBP14*, *PIGV*, *ABCB11*, *ENAM*, *PROP1*, *HPGD*, *SLC26A3*, *SLC19A1*, *AP4B1*, *STAG3*, and *SBF2*.

Figure 8: Cumulative allele frequencies of selected ClinVar variants in 22 AR genes.

Next, we analyzed 248 gene panels from panelApp, containing 1 823 AR disease genes. We identified significant differences (Chi-square test, P-value < 0.01) in cumulative allele frequencies between POL and NFE populations in 30 gene panels (Fig. 9; Supplementary table S6). Among these, “Optic neuropathy” panel was the most enriched in pathogenic variants in POL in comparison to the NFE population.

Figure 9. Cumulative allele frequencies of selected ClinVar variants in AR gene panels.

Selected disease causing variants

In our study, we took a closer look at variants characteristic for the Eastern European or the Slavic populations. We selected three well known diseases, as examples of POL database utility for verification of the genetics of well studied severe diseases.

The most frequent mutation from the list of pathogenic or likely pathogenic variants reported in ClinVar was found in 22 unrelated participants of POL (MAF=1.17%). It was the NG_012655.2:g.12031G>A, p.Trp151Ter variant (rs11555217) located in the *DHCR7* gene which mutations are responsible for Smith-Lemli-Opitz syndrome (SLOS). It is an autosomal recessive disorder characterized by failure to thrive, microcephaly, intellectual disability, cleft palate and multiple birth defects, such as syndactyly of second and third toes, dysmorphic facial features or heart defects (Ryan et al. 1998). Global MAF for the p.Trp151Ter variant is 0.07% (GnomAD v3.1.1) which is significantly lower than the frequency observed in our cohort (Fig. 10A). SLOS occurs as a common recessive disorder in Europe, but is almost absent in most other populations. Previous studies showed that mutations causing this disorder differ among European populations. The founder p.Trp151Ter

variant was the most frequent mutation in Polish SLOS patients compared to patients from other subpopulations such as German or British (Witsch-Baumgartner et al. 2001).

The second interesting example identified in our study is the rare, recessive NM_002485.5:c.657_661del, resulting in p.Lys219fs variant (rs587776650) located in the *NBN* gene which mutations are responsible for Nijmegen breakage syndrome (NBS). NBS is characterized by microcephaly, immunodeficiency and very high predisposition to lymphoid malignancy. It was reported that NBS heterozygosity is also responsible for increased incidence of tumors, especially lymphoma (Seemanová 1990), meaning that the NBS variant contributes to the cancer frequency in a given population. The disease occurs worldwide, but its prevalence was reported to be significantly higher in Poland, Ukraine, Czech Republic and Russia (Kostyuchenko et al. 2009). *NBN* (founder) mutation, 657del5, is estimated to appear in 1 case per 177 newborn (Varon et al. 2000). Global MAF for the NBS causing variant is 0.04% (GnomAD v3.1.1) whereas in our results for the Polish population the frequency of this variant is almost seven fold higher at 0.27% (Fig. 10B). This is the first report of the allele frequency of the NBS causing variant for the Polish population, estimated on the basis of WGS.

Cystic Fibrosis (CF) is another example of an autosomal recessive disease, for which disease-causing allele spectrum was not derived from comprehensive genomic methods in Polish population. The analyses to date have utilized only targeted methods such as INNOLIPA CFTR19 test or Sanger Sequencing of chosen exons (NBS CF working group et al. 2013), without providing data on frequencies of intronic and regulatory site variants. Cystic fibrosis is the most frequent metabolic disease in Poland, with the incidence estimated at 1 per 4394 (NBS CF working group et al. 2013) to 1 per 5000 individuals (Farrell 2008). CF affects mostly lungs and leads to progressive breathing problems and respiratory failure. The disease is caused by mutations located in *CFTR* gene (188 kb, 27 exons), for which more than 2000 variants have been identified to date according to CFTR Mutation Database (data from 28 of June 2021 <http://www.genet.sickkids.on.ca/StatisticsPage.html>). In our study, we observed that the cumulative allele frequency for three pathogenic ClinVar variants found in POL cohort (rs78655421, rs113993960, rs77010898) in the *CFTR* gene is significantly lower than GnomAD NFE population. Specifically, in the case of the most common NM_000492.3(CFTR):c.1521_1523delCTT (p.Phe508del, rs113993960), we observed that the allele frequency in Polish population is slightly lower (0.95%) than in European NFE (1.43%) . However, much lower than previously reported for Polish population, described as almost 3% carrier frequency according to (NBS CF working group et al. 2013; Morral et al. 1994). Surprisingly, in our study, Slavic specific NM_000492.3:c.54-5940_273+10250del (CFTRdele2,3) (Dörk et al. 2000) variant was not found. The reported mutation frequencies for the Polish population for p.Phe508del and CFTRdele2,3 are 57-62% and 1.8-6.2% respectively (Bobadilla et al. 2002; NBS CF working group et al. 2013). The very low frequency of CFTRdele2,3 deletion may be the reason why it was not detected in our cohort of 943 individuals.

Figure 10. Frequency of selected variants in Polish (POL) and gnomAD v.3.1.1 populations: the p.Trp151Ter variant (rs11555217) located in the *DHCR7* gene (A) and the 657del5 variant located in the *NBN* gene (B).

Relationship of the Polish and global populations

In order to explore the diversity of the Polish population and global relationships between POL and other world populations, we performed the principal component analysis (PCA) based on common variants (MAF>1%). In the first comparison with continental populations, we observed that the POL cohort is homogenous and clustered within the European population (Fig. 11A and 11B). After prediction using a random forest method only one sample was located in the AMR population cluster. In PCA of European subpopulations, almost all POL samples (938 out of 943) were clustered with other European ancestries, with 496 individuals belonging to the GBR, 427 to the CEU, 12 to the TSI, and 3 to the IBS subpopulations. Five samples were closer to non-European populations. The results are presented on Fig. 11 (C, D).

Figure 11. Clustering of samples among the World and European subpopulations based on the random forest and PCA prediction. The comparison of (A) the first and second principal components for continental populations; (B) the first and third principal components for continental populations; (C) the first and second principal components for the European subpopulations; (D) the first and third principal components for the European subpopulations.

The above-mentioned results were in accordance with the F_{st} statistics. Considering continental populations, POL was most similar to the European (an average F_{st} of 0.004), and least similar to the African and East Asian populations (average F_{st} of 0.056 and 0.050 respectively). An average F_{st} statistics calculated over sliding windows of 1,000 SNPs is presented on Supplementary Figure S6. Furthermore, we compared the Polish cohort with other European subpopulations from the 1000 Genomes dataset and observed high average F_{st} between the cohorts (Table 2).

Table 2: Fst statistics calculated for Polish cohort vs European subpopulations from the 1000 Genomes project/world dataset.

Population	CEU	FIN	GBR	IBS	TSI
Fst	0.007	0.010	0.007	0.009	0.009

Additionally, density-based spatial clustering among POL individuals (Supplementary Fig. S7) was performed, confirming homogeneity of the cohort with the exception of a few outliers. The average Fst statistics between the POL cluster and the outliers was 0.012, suggesting a non-significant difference between the individuals and the rest of the cohort. The results of all three approaches (PCA, Fst and density-based spatial clustering) agree in demonstrating that POL is highly homogeneous and samples used in the analysis belong to the European populations, both in terms of continental and populational clustering.

We also performed admixture analysis in the unsupervised mode with the 943 unrelated individuals from POL and either the total 1000 Genomes dataset (Supplementary Fig. S8.A) or the European populations (Supplementary Fig. S8.B). Compared against the world populations, the POL cohort was similar to the European populations at low K values ($K = 2$ to 5), and at K above 5 (favoured by the cross-validation analysis for the world dataset) forms a distinct cluster, with some common ancestry with GBR and CEU, and also FIN. In the analysis with only the European populations it formed a distinct cluster that was essentially homogeneous for the values of K less or equal 4. The lower K values were favoured by cross-validation results for the European dataset (the lowest CV error value was for $K = 3$).

Discussion

Although large scale whole genome sequencing projects have arisen worldwide during the last decade, e.g ‘100 000 Genomes’ project in the UK, ‘1+ Million Genomes’ Initiative in Europe or ‘All of Us’ in the United States, country or region-oriented sequencing initiatives were not performed in many areas. Genome of the Netherlands (GoNL) project characterized the genomes of 769 Dutch ancestry individuals (The Genome of the Netherlands Consortium 2014) in 2014. In 2015 the genomes of 2 636 Icelanders were sequenced and described (Gudbjartsson et al. 2015), and in 2016 the population of Sardinia was characterized based on sequencing of 2 120 individuals (Sidore et al. 2015). These country based sequencing analyses significantly improved the SNP imputation performance, confirmed a high degree of geographic clustering of recent mutations and reliably discovered *denovo* mutations (Gudbjartsson et al. 2015; Sidore et al. 2015; The Genome of the Netherlands Consortium 2014), building local genomic resources for clinical genetics and fostering research that was not possible before.

In this study we present results of whole-genome sequencing and comprehensive analyses of a wide spectrum of genomic variation in 1079 individuals of Polish ancestry. We detected and genotyped small and structural variants (SV), runs of homozygosity, mitochondrial haplogroups, and Mendelian inconsistencies, among others. Allele frequencies for 943 unrelated individuals were released as a publicly available resource, The Thousand Polish Genomes (POL) database - the first effort of this scale completed for the Polish population. Two similar initiatives were launched in Poland in recent years, but their results are not openly available to the scientific community. The first one, the Polgenom¹⁰

¹⁰ <https://polgenom.pl>

project, ran in years 2013-2016, collected WGS data from 130 individuals of age 90+ and offered limited paid access to the data. In 2016 a large initiative entitled the Genomic Map of Poland¹¹ was launched, aiming at sequencing a total of 5000 individuals, but its results are not available yet. In general, Central European populations are underrepresented in worldwide genomic databases such as GnomAD, HapMap or 1000 Genomes and the presented work fills this gap, and equips researchers with large and freely available reference dataset of variant frequencies.

The sequencing data gives a unique opportunity to interrogate rare variation. The rare alleles are considered relatively new in the population (Veltman and Brunner 2012) and they tend to cluster geographically more than the common alleles (Gravel et al. 2011; Mathieson and McVean 2012; The 1000 Genomes Project Consortium 2015). In the first phase of the POL project, which is described in the presented article, we focused on the identification of rare diseases-related variants, which may appear more frequently in Polish population in comparison to other European populations. We identified 736 pathogenic variants according to ClinVar, but more than 70% of them were present in a single person only, hindering reliable frequency estimates on a cohort of this size. The allele frequencies of the remaining pathogenic variants were up to 1.17 % (22/1886), which may have visible clinical consequences in the population. Among 16 pathogenic variants in the ACMG actionable genes identified in POL only 1 variant reached the frequency of 0.15 % (3/1886) (rs1217805587 in *BRCA1* gene). The overall incidence of secondary findings in Polish population matched the ACMG estimations. We suggest that national guidelines, available in Polish language, should be issued to address the importance of genetic counselling for clinical WES/WGS results concerning incidental findings specific to our population, as availability to whole genome analyses will successively increase.

The analysis of OMIM AR disease genes let us identify as many as 22 genes (Supplementary Table S5) with significantly different cumulative allele frequencies in POL vs European NFE populations. Variants of most of these genes are associated with the founder's effect in the central european population and have shown the most striking differences between the POL population and the rest of Europe. The whole set of 22 genes may be taken into consideration by researchers and clinicians interested in specific rare diseases and following their incidence in the population.

Undoubtedly, apart of small variants, a valuable part of the released data is the structural variant (SV) dataset, called jointly on all unrelated samples in the cohort. Historically, SVs have been understudied due to methodological limitations. However, they collectively affect a larger part of the genome than the small variants (Sudmant et al. 2015), which we also observed in this cohort. Compared to other large-scale studies of SVs (Chiang et al. 2017; Collins et al. 2020; Hehir-Kwa et al. 2016), we report lower SV counts per individual, and larger fractions of affected genomic sequence, which may stem from methodological differences that are known to considerably affect structural variant detection (Cameron et al., 2019). We chose a single-tool approach with joint genotyping of the SVs, which could have resulted in higher false-positive rates and lower sensitivity than ensemble methods used in other studies (Collins et al., 2020; Sudmant et al., 2015). However, considering the uniqueness of this data and the role it could play, for instance for rare-variant filtering, we decided to release the unfiltered SV frequencies. In future releases of the database, the spurious SV calls will be marked, allowing researchers to more precisely compare SV characteristics between cohorts. We also note that further analyses combined with populational aCGH or optical mapping data are needed to explore and validate the diversity and clinical utility of SV patterns in Polish population and broader context as their pathogenic potential and involvement in complex genetic traits is still poorly understood.

It was previously described that the Polish population is relatively homogenous, with a low level of genetic diversity (Jarczak et al. 2019). Small genetic differences are present at regional level and are in correlation to patterns of migrations in the past (Jarczak et al. 2019). In the current study the analysis of mtDNA haplogroups was performed on 943 unrelated individuals. For the majority of people in our

¹¹ <https://www.genompolski.pl>

cohort (44 %) the haplogroup H was assigned, which is consistent with previous results for Polish and Slavic populations (Grzybowski et al. 2007; Jarczak et al. 2019; Mielnik-Sikorska et al. 2013). Other studies on mitochondrial DNA show that Poles as a population are characterized by different European haplogroups, with dominance of West Eurasian, Central and Eastern European haplogroups (Jarczak et al. 2019; B. A. Malyarchuk et al. 2002; Mielnik-Sikorska et al. 2013). It was also shown that Polish population is almost indistinguishable from other European nations, except for the sub-haplogroups U4a and HV3a which are predominantly found in Poles and Russians (B. Malyarchuk et al. 2008; B. A. Malyarchuk et al. 2002). In this project we confirmed only the presence of the U4a haplogroup (28 individuals), which is assumed to be of central-european origin (B. A. Malyarchuk et al. 2002).

The analyses of Polish population similarities to other populations performed in presented work clearly proved the homogeneity of POL as well as the closest ancestry with GBR and CEU populations. The admixture analysis also confirmed that the Polish cohort is closest to the other European populations in the 1000 genomes dataset. The Polish cohort forms a homogeneous cluster that is distinct from the other European populations, again confirming its low level of genetic differentiation. Due to the lack of other samples from Central and Eastern Europe in this dataset, making any detailed inference about the ancestry of the Polish population compared to the rest of the continent is impossible at the moment. In comparison to populations from northwestern and southern Europe that are present in the 1000 Genomes dataset, the Polish population is expected to have more traces of the steppe ancestry (Lazaridis, 2018), but the populations available in the dataset are not sufficient to test this hypothesis. One cannot also exclude the possibility that the distinct clustering of the Polish cohort is related to the differences in WGS data processing methods used to obtain different datasets.

Runs of homozygosity (ROH) analysis identifies the stretches of contiguous homozygous sites in an individual and are used to measure the level of inbreeding and recessive inheritance. The longer the total length of ROH in individuals the closer the inbreeding in a population, which increases the risk of recessive disease and decreases reproductive fitness of the offspring (Hoffman et al. 2014). The Polish population characterised in this study perfectly matches the European ROH characteristics). This suggests that the inbreeding level and potential risk of recessive diseases caused by homozygosity is relatively low. There is however some risk that sampling bias could have influenced the ROH analysis as most of the volunteering participants were city inhabitants. On the contrary ROH density increases in rural areas, less mobile and more prone to close-kin marriages. Unfortunately, this type of bias occurs in most large scale genomic studies to date (Dean et al. 2017).

As with many populational, volunteer-based projects, our data are not bias free. Our cohort was primarily recruited to perform the GWAS study on COVID-19 susceptibility, therefore we didn't focus on collecting samples from Polish minorities like Kaschubians, Karaims, Lemkos or Polish Jews. We have also not corrected the much higher proportion of donors recruited from municipal areas compared to rural, as recruitment of volunteers from small villages is a cumbersome, costly process due to lack of trust and no means of effective project promotion in these communities. Also, the proportion of individuals from different voivodeships is skewed in favor of Masovia and Greater Poland, probably because of more intense promotion of the project that was conducted there. As a result, the far-eastern voivodeships were underrepresented which could have led to missing variants that could have been native e.g. to Tatar or Russian populations. Moreover, the cohort of more than 1000 genomes, however impressive, does not enable to reliably estimate the prevalence of variants less common than 0,5%, which could be very useful in clinical setting to re-evaluate and reclassify many VUS variants (variant of uncertain significance) to benign site of the spectrum.

Although in terms of sample size our project does not compare to the world's largest, it remains one of the largest open allele-frequency datasets generated on high-coverage WGS data and the largest of Slavic population. Results presented in this paper are only the tip of an iceberg and should be taken as an invitation for further explorations. Our main goal was to generate an open-source dataset available

for all researchers around the world, thus, filling in an existing gap in world's datasets so far lacking Eastern-European samples. We believe that such collaborative efforts will be soon performed in many countries delivering excellent results for future research of many disciplines.

Acknowledgements

Authors would like to thank the medical personnel of Central Clinical Hospital of Ministry of the Interior and Administration in Warsaw, Multidisciplinary Municipal Hospital of Józef Struś in Poznań and ALAB Laboratoria Sp.z o.o. as well as SiePomaga Foundation for collaboration and blood sample collection.

This work was partially supported by the Polish National Science Centre grant No. SZPITALE-JEDNOIMIENNE/2/2020 and by the Medical Research Agency grant No 2020/ABM /COVID19/0022

References

- ACMG Secondary Findings Working Group, David T. Miller, Kristy Lee, Wendy K. Chung, Adam S. Gordon, Gail E. Herman, Teri E. Klein, et al. 2021. “ACMG SF v3.0 List for Reporting of Secondary Findings in Clinical Exome and Genome Sequencing: A Policy Statement of the American College of Medical Genetics and Genomics (ACMG).” *Genetics in Medicine*, May. <https://doi.org/10.1038/s41436-021-01172-3>.
- Adzhubei, Ivan A., Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. 2010. “A Method and Server for Predicting Damaging Missense Mutations.” *Nature Methods* 7 (4): 248–49. <https://doi.org/10.1038/nmeth0410-248>.
- Alexander, David H., John Novembre, and Kenneth Lange. 2009. “Fast Model-Based Estimation of Ancestry in Unrelated Individuals.” *Genome Research* 19 (9): 1655–64. <https://doi.org/10.1101/gr.094052.109>.
- Altshuler, D., M. J. Daly, and E. S. Lander. 2008. “Genetic Mapping in Human Disease.” *Science* 322 (5903): 881–88. <https://doi.org/10.1126/science.1156409>.
- Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74. <https://doi.org/10.1038/nature15393>.
- Black, Jennifer O. 2016. “Xeroderma Pigmentosum.” *Head and Neck Pathology* 10 (2): 139–44. <https://doi.org/10.1007/s12105-016-0707-8>.
- Bobadilla, Joseph L., Milan Macek, Jason P. Fine, and Philip M. Farrell. 2002. “Cystic Fibrosis: A Worldwide Analysis of CFTR Mutations--Correlation with Incidence Data and Application to Screening.” *Human Mutation* 19 (6): 575–606. <https://doi.org/10.1002/humu.10041>.
- Bochukova, Elena G., Ni Huang, Julia Keogh, Elana Henning, Carolin Purmann, Kasia Blaszczyk, Sadia Saeed, et al. 2010. “Large, Rare Chromosomal Deletions Associated with Severe Early-Onset Obesity.” *Nature* 463 (7281): 666–70. <https://doi.org/10.1038/nature08689>.
- Cameron, Daniel L., Leon Di Stefano, and Anthony T. Papenfuss. 2019. “Comprehensive Evaluation and Characterisation of Short Read General-Purpose Structural Variant Calling Software.” *Nature Communications* 10 (1): 3240. <https://doi.org/10.1038/s41467-019-11146-4>.
- Chiang, Colby, Ryan M. Layer, Gregory G. Faust, Michael R. Lindberg, David B. Rose, Erik P. Garrison, Gabor T. Marth, Aaron R. Quinlan, and Ira M. Hall. 2015. “SpeedSeq: Ultra-Fast Personal Genome Analysis and Interpretation.” *Nature Methods* 12 (10): 966–68. <https://doi.org/10.1038/nmeth.3505>.
- Chiang, Colby, Alexandra J. Scott, Joe R. Davis, Emily K. Tsang, Xin Li, Yungil Kim, Tarik Hadzic, et al. 2017. “The Impact of Structural Variation on Human Gene Expression.” *Nature Genetics* 49 (5): 692–99. <https://doi.org/10.1038/ng.3834>.
- Collins, Ryan L., Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C. Francioli, Amit V. Khera, et al. 2020. “A Structural Variation Reference for Medical and Population Genetics.” *Nature* 581 (7809): 444–51. <https://doi.org/10.1038/s41586-020-2287-8>.
- Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. 2021. “Twelve Years of SAMtools and BCFtools.” *GigaScience* 10 (2). <https://doi.org/10.1093/gigascience/giab008>.
- Dean, Caress, Amanda J. Fogleman, Whitney E. Zahnd, Alexander E. Lipka, Ripan Singh Malhi, Kristin R. Delfino, and Wiley D. Jenkins. 2017. “Engaging Rural Communities in Genetic Research: Challenges and Opportunities.” *Journal of Community Genetics* 8 (3): 209–19. <https://doi.org/10.1007/s12687-017-0304-x>.
- Donnelly, Peter. 2008. “Progress and Challenges in Genome-Wide Association Studies in Humans.” *Nature* 456 (7223): 728–31. <https://doi.org/10.1038/nature07631>.
- Dörk, T., M. Macek, F. Mekus, B. Tümmler, J. Tzountzouris, T. Casals, A. Krebsová, et al. 2000. “Characterization of a Novel 21-Kb Deletion, CFTRdele2,3(21 Kb), in the CFTR Gene: A

- Cystic Fibrosis Mutation of Slavic Origin Common in Central and East Europe.” *Human Genetics* 106 (3): 259–68. <https://doi.org/10.1007/s004390000246>.
- Exome Aggregation Consortium, Monkol Lek, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, et al. 2016. “Analysis of Protein-Coding Genetic Variation in 60,706 Humans.” *Nature* 536 (7616): 285–91. <https://doi.org/10.1038/nature19057>.
- Farrell, Philip M. 2008. “The Prevalence of Cystic Fibrosis in the European Union.” *Journal of Cystic Fibrosis* 7 (5): 450–53. <https://doi.org/10.1016/j.jcf.2008.03.007>.
- Faust, Gregory G., and Ira M. Hall. 2014. “SAMBLASTER: Fast Duplicate Marking and Structural Variant Read Extraction.” *Bioinformatics* 30 (17): 2503–5. <https://doi.org/10.1093/bioinformatics/btu314>.
- Frazer, Kelly A., Sarah S. Murray, Nicholas J. Schork, and Eric J. Topol. 2009. “Human Genetic Variation and Its Contribution to Complex Traits.” *Nature Reviews Genetics* 10 (4): 241–51. <https://doi.org/10.1038/nrg2554>.
- Garrison, Erik, and Gabor Marth. 2012. “Haplotype-Based Variant Detection from Short-Read Sequencing.” *ArXiv:1207.3907 [q-Bio]*, July. <http://arxiv.org/abs/1207.3907>.
- Goodstadt, Leo. 2010. “Ruffus: A Lightweight Python Library for Computational Pipelines.” *Bioinformatics* 26 (21): 2778–79. <https://doi.org/10.1093/bioinformatics/btq524>.
- Gravel, S., B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, et al. 2011. “Demographic History and Rare Allele Sharing among Human Populations.” *Proceedings of the National Academy of Sciences* 108 (29): 11983–88. <https://doi.org/10.1073/pnas.1019276108>.
- Grzybowski, Tomasz, Boris A. Malyarchuk, Miroslava V. Derenko, Maria A. Perkova, Jarosław Bednarek, and Marcin Woźniak. 2007. “Complex Interactions of the Eastern and Western Slavic Populations with Other European Groups as Revealed by Mitochondrial DNA Analysis.” *Forensic Science International: Genetics* 1 (2): 141–47. <https://doi.org/10.1016/j.fsigen.2007.01.010>.
- Gudbjartsson, Daniel F., Hannes Helgason, Sigurjon A. Gudjonsson, Florian Zink, Asmundur Oddson, Arnaldur Gylfason, Soren Besenbacher, et al. 2015. “Large-Scale Whole-Genome Sequencing of the Icelandic Population.” *Nature Genetics* 47 (5): 435–44. <https://doi.org/10.1038/ng.3247>.
- Hehir-Kwa, Jayne Y., Tobias Marschall, Wigard P. Kloosterman, Laurent C. Francioli, Jasmijn A. Baaijens, Louis J. Dijkstra, Abdel Abdellaoui, et al. 2016. “A High-Quality Human Reference Panel Reveals the Complexity and Distribution of Genomic Structural Variants.” *Nature Communications* 7 (October): 12989. <https://doi.org/10.1038/ncomms12989>.
- Hindorf, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. 2009. “Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits.” *Proceedings of the National Academy of Sciences* 106 (23): 9362–67. <https://doi.org/10.1073/pnas.0903103106>.
- Hinds, D. A. 2005. “Whole-Genome Patterns of Common DNA Variation in Three Human Populations.” *Science* 307 (5712): 1072–79. <https://doi.org/10.1126/science.1105436>.
- Hoffman, Joseph I., Fraser Simpson, Patrice David, Jolianne M. Rijks, Thijs Kuiken, Michael A. S. Thorne, Robert C. Lacy, and Kanchon K. Dasmahapatra. 2014. “High-Throughput Sequencing Reveals Inbreeding Depression in a Natural Population.” *Proceedings of the National Academy of Sciences* 111 (10): 3775–80.
- International HapMap Consortium, Kelly A. Frazer, Dennis G. Ballinger, David R. Cox, David A. Hinds, Laura L. Stuve, Richard A. Gibbs, et al. 2007. “A Second Generation Human Haplotype Map of over 3.1 Million SNPs.” *Nature* 449 (7164): 851–61. <https://doi.org/10.1038/nature06258>.
- Jarczak, Justyna, Łukasz Grochowalski, Błażej Marciniak, Jakub Lach, Marcin Słomka, Marta Sobalska-Kwapis, Wiesław Lorkiewicz, Łukasz Pułaski, and Dominik Strapagiel. 2019. “Mitochondrial DNA Variability of the Polish Population.” *European Journal of Human Genetics* 27 (8): 1304–14. <https://doi.org/10.1038/s41431-019-0381-x>.
- Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. “The Mutational Constraint Spectrum Quantified

- from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43.
<https://doi.org/10.1038/s41586-020-2308-7>.
- Kostyuchenko, Larysa, Halyna Makuch, Nataliya Kitsera, Romana Polishchuk, Nataliya Markevych, and Hajane Akopian. 2009. “Clinical Immunology
Nijmegen Breakage Syndrome in Ukraine: Diagnostics and Follow-Up.” *Central European Journal of Immunology* 34 (1): 46–52.
- Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2018. “ClinVar: Improving Access to Variant Interpretations and Supporting Evidence.” *Nucleic Acids Research* 46 (D1): D1062–67.
<https://doi.org/10.1093/nar/gkx1153>.
- Lazaridis, Iosif. 2018. “The Evolutionary History of Human Populations in Europe.” *Current Opinion in Genetics & Development* 53 (December): 21–27.
<https://doi.org/10.1016/j.gde.2018.06.007>.
- Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *ArXiv* 1303 (March).
- Li, Yingrui, Nicolas Vinckenbosch, Geng Tian, Emilia Huerta-Sanchez, Tao Jiang, Hui Jiang, Anders Albrechtsen, et al. 2010. “Resequencing of 200 Human Exomes Identifies an Excess of Low-Frequency Non-Synonymous Coding Variants.” *Nature Genetics* 42 (11): 969–72.
<https://doi.org/10.1038/ng.680>.
- Malyarchuk, B. A., T. Grzybowski, M. V. Derenko, J. Czarny, M. Wozniak, and D. Miscicka-Sliwka. 2002. “Mitochondrial DNA Variability in Poles and Russians.” *Annals of Human Genetics* 66 (4): 261–83. <https://doi.org/10.1046/j.1469-1809.2002.00116.x>.
- Malyarchuk, B., T. Grzybowski, M. Derenko, M. Perkova, T. Vanecek, J. Lazur, P. Gomolcak, and I. Tsybovsky. 2008. “Mitochondrial DNA Phylogeny in Eastern and Western Slavs.” *Molecular Biology and Evolution* 25 (8): 1651–58. <https://doi.org/10.1093/molbev/msn114>.
- Manolio, Teri A. 2013. “Bringing Genome-Wide Association Findings into Clinical Use.” *Nature Reviews Genetics* 14 (8): 549–58. <https://doi.org/10.1038/nrg3523>.
- Manolio, Teri A., and Francis S. Collins. 2009. “The HapMap and Genome-Wide Association Studies in Diagnosis and Therapy.” *Annual Review of Medicine* 60 (1): 443–56.
<https://doi.org/10.1146/annurev.med.60.061907.093117>.
- Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. “Finding the Missing Heritability of Complex Diseases.” *Nature* 461 (7265): 747–53. <https://doi.org/10.1038/nature08494>.
- Mathieson, Iain, and Gil McVean. 2012. “Differential Confounding of Rare and Common Variants in Spatially Structured Populations.” *Nature Genetics* 44 (3): 243–46.
<https://doi.org/10.1038/ng.1074>.
- McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. “The Ensembl Variant Effect Predictor.” *Genome Biology* 17 (1): 122–122. <https://doi.org/10.1186/s13059-016-0974-4>.
- Mielnik-Sikorska, Marta, Patrycja Daga, Boris Malyarchuk, Miroslava Derenko, Katarzyna Skonieczna, Maria Perkova, Tadeusz Dobosz, and Tomasz Grzybowski. 2013. “The History of Slavs Inferred from Complete Mitochondrial Genome Sequences.” Edited by Luísa Maria Sousa Mesquita Pereira. *PLoS ONE* 8 (1): e54360.
<https://doi.org/10.1371/journal.pone.0054360>.
- Morral, N., J. Bertranpetit, X. Estivill, V. Nunes, T. Casals, J. Giménez, A. Reis, et al. 1994. “The Origin of the Major Cystic Fibrosis Mutation ($\Delta F508$) in European Populations.” *Nature Genetics* 7 (2): 169–75. <https://doi.org/10.1038/ng0694-169>.
- Musich, Phillip R., Zhengke Li, and Yue Zou. 2017. “Xeroderma Pigmentosa Group A (XPA), Nucleotide Excision Repair and Regulation by ATR in Response to Ultraviolet Irradiation.” In *Ultraviolet Light in Human Health, Diseases and Environment*, edited by Shamim I. Ahmad, 41–54. Advances in Experimental Medicine and Biology. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-56017-5_4.
- Narasimhan, Vagheesh, Petr Danecek, Aylwyn Scally, Yali Xue, Chris Tyler-Smith, and Richard Durbin. 2016. “BCFtools/RoH: A Hidden Markov Model Approach for Detecting Autozygosity from next-Generation Sequencing Data.” *Bioinformatics* 32 (11): 1749–51.

- <https://doi.org/10.1093/bioinformatics/btw044>.
- NBS CF working group, Agnieszka Sobczykńska-Tomaszewska, Mariusz Ołtarzewski, Kamila Czerska, Katarzyna Wertheim-Tysarowska, Dorota Sands, Jarosław Walkowiak, Jerzy Bal, and Tadeusz Mazurczak. 2013. “Newborn Screening for Cystic Fibrosis: Polish 4 Years’ Experience with CFTR Sequencing Strategy.” *European Journal of Human Genetics* 21 (4): 391–96. <https://doi.org/10.1038/ejhg.2012.180>.
- Nejentsev, S., N. Walker, D. Riches, M. Egholm, and J. A. Todd. 2009. “Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes.” *Science* 324 (5925): 387–89. <https://doi.org/10.1126/science.1167728>.
- NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, Daniel Taliun, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, et al. 2021. “Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program.” *Nature* 590 (7845): 290–99. <https://doi.org/10.1038/s41586-021-03205-y>.
- Pedersen, Brent, and Aaron Quinlan. 2017. *Mosdepth: Quick Coverage Calculation for Genomes and Exomes*. <https://doi.org/10.1101/185843>.
- Poplin, Ryan, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, et al. 2018. “A Universal SNP and Small-Indel Variant Caller Using Deep Neural Networks.” *Nature Biotechnology* 36 (10): 983–87. <https://doi.org/10.1038/nbt.4235>.
- Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses.” *American Journal of Human Genetics* 81 (3): 559–75.
- Quang, Daniel, Yifei Chen, and Xiaohui Xie. 2015. “DANN: A Deep Learning Approach for Annotating the Pathogenicity of Genetic Variants.” *Bioinformatics* 31 (5): 761–63. <https://doi.org/10.1093/bioinformatics/btu703>.
- Quinodoz, Mathieu, Virginie G. Peter, Nicola Bedoni, Bérlyl Royer Bertrand, Katarina Cisarova, Arash Salmaninejad, Neda Sepahi, et al. 2021. “AutoMap Is a High Performance Homozygosity Mapping Tool Using Next-Generation Sequencing Data.” *Nature Communications* 12 (1): 518. <https://doi.org/10.1038/s41467-020-20584-4>.
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Redon, Richard, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, et al. 2006. “Global Variation in Copy Number in the Human Genome.” *Nature* 444 (7118): 444–54. <https://doi.org/10.1038/nature05329>.
- Ryan, A. K., K. Bartlett, P. Clayton, S. Eaton, L. Mills, D. Donnai, R. M. Winter, and J. Burn. 1998. “Smith-Lemli-Opitz Syndrome: A Variable Clinical and Biochemical Phenotype.” *Journal of Medical Genetics* 35 (7): 558–65. <https://doi.org/10.1136/jmg.35.7.558>.
- Seemanová, E. 1990. “An Increased Risk for Malignant Neoplasms in Heterozygotes for a Syndrome of Microcephaly, Normal Intelligence, Growth Retardation, Remarkable Facies, Immunodeficiency and Chromosomal Instability.” *Mutation Research* 238 (3): 321–24. [https://doi.org/10.1016/0165-1110\(90\)90024-6](https://doi.org/10.1016/0165-1110(90)90024-6).
- Sidore, Carlo, Fabio Busonero, Andrea Maschio, Eleonora Porcu, Silvia Naitza, Magdalena Zoledziwska, Antonella Mulas, et al. 2015. “Genome Sequencing Elucidates Sardinian Genetic Architecture and Augments Association Analyses for Lipid and Blood Inflammatory Markers.” *Nature Genetics* 47 (11): 1272–81. <https://doi.org/10.1038/ng.3368>.
- Sim, Ngak-Leng, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C. Ng. 2012. “SIFT Web Server: Predicting Effects of Amino Acid Substitutions on Proteins.” *Nucleic Acids Research* 40 (W1): W452–57. <https://doi.org/10.1093/nar/gks539>.
- Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. “An Integrated Map of Structural Variation in 2,504 Human Genomes.” *Nature* 526 (7571): 75–81. <https://doi.org/10.1038/nature15394>.
- Tange, O. 2011. “GNU Parallel: The Command-Line Power Tool.” 2011. <https://www.usenix.org/publications/login/february-2011-volume-36-number-1/gnu-parallel-command-line-power-tool>.
- Tarasov, Artem, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. 2015. “Sambamba: Fast Processing of NGS Alignment Formats.” *Bioinformatics (Oxford, England)* 31 (12):

- 2032–34. <https://doi.org/10.1093/bioinformatics/btv098>.
- The 1000 Genomes Project Consortium. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74. <https://doi.org/10.1038/nature15393>.
- The Genome of the Netherlands Consortium. 2014. “Whole-Genome Sequence Variation, Population Structure and Demographic History of the Dutch Population.” *Nature Genetics* 46 (8): 818–25. <https://doi.org/10.1038/ng.3021>.
- The International HapMap 3 Consortium. 2010. “Integrating Common and Rare Genetic Variation in Diverse Human Populations.” *Nature* 467 (7311): 52–58. <https://doi.org/10.1038/nature09298>.
- †The International HapMap Consortium. 2003. “The International HapMap Project.” *Nature* 426 (6968): 789–96. <https://doi.org/10.1038/nature02168>.
- The Wellcome Trust Case Control Consortium, Donald F. Conrad, Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, et al. 2010. “Origins and Functional Impact of Copy Number Variation in the Human Genome.” *Nature* 464 (7289): 704–12. <https://doi.org/10.1038/nature08516>.
- Varon, R., E. Seemanova, K. Chrzanowska, O. Hnateyko, D. Piekutowska-Abramczuk, M. Krajewska-Walasek, J. Sykut-Cegielska, K. Sperling, and A. Reis. 2000. “Clinical Ascertainment of Nijmegen Breakage Syndrome (NBS) and Prevalence of the Major Mutation, 657del5, in Three Slav Populations.” *European Journal of Human Genetics: EJHG* 8 (11): 900–902. <https://doi.org/10.1038/sj.ejhg.5200554>.
- Veltman, Joris A., and Han G. Brunner. 2012. “De Novo Mutations in Human Genetic Disease.” *Nature Reviews Genetics* 13 (8): 565–75. <https://doi.org/10.1038/nrg3241>.
- Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. “Five Years of GWAS Discovery.” *The American Journal of Human Genetics* 90 (1): 7–24. <https://doi.org/10.1016/j.ajhg.2011.11.029>.
- Weir, B. S., and C. Clark Cockerham. 1984. “Estimating F-Statistics for the Analysis of Population Structure.” *Evolution* 38 (6): 1358–70. <https://doi.org/10.2307/2408641>.
- Weischenfeldt, Joachim, Orsolya Symmons, François Spitz, and Jan O. Korbel. 2013. “Phenotypic Impact of Genomic Structural Variation: Insights from and for Human Disease.” *Nature Reviews Genetics* 14 (2): 125–38. <https://doi.org/10.1038/nrg3373>.
- Weissensteiner, Hansi, Dominic Pacher, Anita Kloss-Brandstätter, Lukas Forer, Günther Specht, Hans-Jürgen Bandelt, Florian Kronenberg, Antonio Salas, and Sebastian Schönherr. 2016. “HaploGrep 2: Mitochondrial Haplogroup Classification in the Era of High-Throughput Sequencing.” *Nucleic Acids Research* 44 (W1): W58–63. <https://doi.org/10.1093/nar/gkw233>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemond, et al. 2019. “Welcome to the Tidyverse.” *Journal of Open Source Software* 4 (November): 1686. <https://doi.org/10.21105/joss.01686>.
- Witsch-Baumgartner, M., E. Ciara, J. Löffler, H. J. Menzel, U. Seedorf, J. Burn, G. Gillesen-Kaesbach, et al. 2001. “Frequency Gradients of DHCR7 Mutations in Patients with Smith-Lemli-Opitz Syndrome in Europe: Evidence for Different Origins of Common Mutations.” *European Journal of Human Genetics* 9 (1): 45–50. <https://doi.org/10.1038/sj.ejhg.5200579>.
- Yun, Taedong, Helen Li, Pi-Chuan Chang, Michael F. Lin, Andrew Carroll, and Cory Y. McLean. 2021. “Accurate, Scalable Cohort Variant Calls Using DeepVariant and GLnexus.” *Bioinformatics (Oxford, England)*, January, btaa1081. <https://doi.org/10.1093/bioinformatics/btaa1081>.