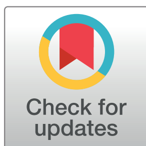
RESEARCH ARTICLE

The sequencing and interpretation of the genome obtained from a Serbian individual

Wazim Mohammed Ismail[1☯], **Kymberleigh A. Pagel**[1☯], **Vikas Pejaver**[1], **Simo V. Zhang**[1], **Sofia Casasa**[2], **Matthew Mort**[3], **David N. Cooper**[3], **Matthew W. Hahn**[1,2], **Predrag Radivojac**[4]*

1 Department of Computer Science, Indiana University, Bloomington, Indiana, United States of America, **2** Department of Biology, Indiana University, Bloomington, Indiana, United States of America, **3** Institute of Medical Genetics, Cardiff University, Cardiff, United Kingdom, **4** College of Computer and Information Science, Northeastern University, Boston, Massachusetts, United States of America

☯ These authors contributed equally to this work.
* predrag@northeastern.edu



OPEN ACCESS

Citation: Mohammed Ismail W, Pagel KA, Pejaver V, Zhang SV, Casasa S, Mort M, et al. (2018) The sequencing and interpretation of the genome obtained from a Serbian individual. PLoS ONE 13 (12): e0208901. https://doi.org/10.1371/journal.pone.0208901

Editor: Marc Robinson-Rechavi, University of Lausanne, SWITZERLAND

Received: April 24, 2018

Accepted: November 26, 2018

Published: December 19, 2018

Copyright: © 2018 Mohammed Ismail et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The raw data files are freely available at https://my.pgp-hms.org/profile/hu3BDC4B.

Funding: This work was supported in part by the Center for Bioinformatics Research at Indiana University.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Recent genetic studies and whole-genome sequencing projects have greatly improved our understanding of human variation and clinically actionable genetic information. Smaller ethnic populations, however, remain underrepresented in both individual and large-scale sequencing efforts and hence present an opportunity to discover new variants of biomedical and demographic significance. This report describes the sequencing and analysis of a genome obtained from an individual of Serbian origin, introducing tens of thousands of previously unknown variants to the currently available pool. Ancestry analysis places this individual in close proximity to Central and Eastern European populations; i.e., closest to Croatian, Bulgarian and Hungarian individuals and, in terms of other Europeans, furthest from Ashkenazi Jewish, Spanish, Sicilian and Baltic individuals. Our analysis confirmed gene flow between Neanderthal and ancestral pan-European populations, with similar contributions to the Serbian genome as those observed in other European groups. Finally, to assess the burden of potentially disease-causing/clinically relevant variation in the sequenced genome, we utilized manually curated genotype-phenotype association databases and variant-effect predictors. We identified several variants that have previously been associated with severe early-onset disease that is not evident in the proband, as well as putatively impactful variants that could yet prove to be clinically relevant to the proband over the next decades. The presence of numerous private and low-frequency variants, along with the observed and predicted disease-causing mutations in this genome, exemplify some of the global challenges of genome interpretation, especially in the context of under-studied ethnic groups.

Introduction

The genetic variation between individuals accounts for much of observed human diversity and has the potential to provide information on phenotypic outcomes of clinical consequence. Studies of genetic variation provided by individual genome sequences have revealed that this

variation differs both within and between populations, and also varies considerably depending upon the population [1]. Moreover, characterization of genetic variation of individuals from multiple populations has revealed a correlation between genetic and geographic distances, and has become relevant for determining genetic ancestry and geographic origin [2–6]. Therefore, the characterization of genetic variation has been of major interest for diverse research fields, including medical, biological and anthropological sciences [2–10].

Sequencing of the first human genomes revealed that most genetic variation is derived from single nucleotide variants (SNVs), although insertions and deletions (indels) account for the majority of the variant nucleotides [11]. The increased accessibility of DNA sequencing has contributed to individual efforts from a range of distinct populations. To date, individual genomes from American [11, 12], Han Chinese [13], Russian [14], Khoisan [15], Bantu [15], Japanese [16], German [17], Gujarati Indian [18], Estonian [19], Pakistani [20] and Mongolian [21] populations have been sequenced and analyzed, among many others [1].

Larger-scale efforts to characterize human genetic variation have demonstrated that individuals from different populations carry particular combinations of rare and low-frequency variants. The 1000 Genomes Project Consortium has estimated that 86% of all variants are confined to a single continental group and that about 10% of variants observed in a population are private to that population [1]. Population-specific variants have the potential to be of both functional and biomedical importance [7, 22–24]. Furthermore, evidence of biologically meaningful population-specific variation [25] emphasizes the need for ethnically relevant reference genomes, as has been performed, for example, for the Korean population [26]. Although we are not claiming to have introduced a new reference genome here, it is nevertheless important to expand our sequencing efforts across diverse populations, particularly those that have not been previously studied [10, 27].

In this paper, we describe the sequencing of the first genome of an individual of Serbian origin, a member of a relatively small population in Central to Southeastern Europe. We identify tens of thousands of novel genetic variants in this individual, more than a hundred of which map to protein-coding regions and several hundred of which reside in close proximity to gene coding regions. The extent of observed genetic variation allowed comparisons with extant European populations and reaffirms support for the hypothesis of close correspondence between genetic and geographic distances [2]. These results contribute to ongoing efforts to understand human genetic variation and its geographic distribution, as well as placing the Serbian genome within the context of the broader European population structure. Testing for Neanderthal introgression in the genome, we find evidence to suggest gene flow from Neanderthal to an ancestral pan-European genome, with the Serbian genome being placed within the range of other European populations. After variant annotation, we assess the burden of potentially pathogenic variation present in this genome and identify variants of putative clinical and pharmacogenetic relevance. Finally, we draw conclusions pertaining to the phenotypic consequences and biomedical interpretation of individually sequenced genomes.

Materials and methods

Donor information

The individual whose genome was sequenced and analyzed is a male of Serbian descent. The data, both derived and raw, are publicly available through the Personal Genome Project website [28], participant ID: hu3BDC4B.

Sample collection and DNA sequencing

Two milliliters of saliva were self-collected by the donor and stored using the DNA Genotek Oragene DISCOVER (OGR-500) sample collection kit. Extraction of DNA from the sample and subsequent sequencing were performed at the BGI (Shenzhen, China) on an Illumina HiSeq 2000 sequencer, using standard protocols. To minimize the likelihood of systematic bias in sampling, two libraries were prepared with an insert size of 500 bp each, with paired-end reads of length 90 bp. Sequencing was then carried out in four lanes for each library to ensure at least 30-fold coverage.

Read mapping and variant calling

Single Nucleotide Variants (SNVs) and indels were called using four different pipelines through a combination of two read mappers and two variant callers. The GRCh37 human genome was used as the reference genome to map the paired-end reads. The two read mappers used were BWA-MEM [29] and Bowtie2 [30]. The two variant callers were GATK [31] and Platypus [32]. The GATK pipeline included additional read and variant processing steps such as duplicate removal using Picard tools [33], base quality score recalibration, indel realignment, and genotyping and variant quality score recalibration using GATK, all used according to GATK best practice recommendations [34, 35].

As described later in the Results, variants identified using the BWA+ GATK pipeline were used for all downstream analysis. Variants in the intersection of all four pipelines (two read mappers and two variant callers) were considered to be confidently identified, where the intersection is defined as variant calls for which the chromosome, position, reference, and alternate fields in the VCF files were identical. All variant calls were subsequently annotated with information from NCBI RefSeq using ANNOVAR [36]. We estimated the amount of novel variation expected to be observed from the first individual in a previously uncharacterized population utilizing the 1000 Genomes Project Phase 3 VCF files [37]. To do this, we carried out a leave-one-population-out procedure; i.e., we excluded one of the 26 populations at a time and for each individual in the excluded population, calculated the fraction of variants not seen in any of the individuals from the remaining 25 populations. The calculated fractions of novel variants were used to understand the expected novelty when sequencing an individual from a new population, given a sample of a particular size of previously sequenced individuals from different populations.

Structural variants (SVs) were called using Structural Variation Engine (SVE) and FusorSV [38]. SVE is an execution engine for an ensemble of SV calling algorithms containing BreakDancer [39], BreakSeq2 [40], cnMOPS [41], CNVnator [42], DELLY [43], GenomeSTRiP [44, 45], Hydra [46], and LUMPY [47]. The Docker image of SVE was used to run all the stages with default parameters. All but GenomeSTRiP completed without errors. The Docker image of FusorSV was then used to merge the results from the remaining seven SV callers, using the default fusion model. SVint [48] was used to subsequently annotate the structural variants. Scripts and documentation for parameters used to run all the pipelines described in this study were added to the Personal Genome Project website, participant ID hu3BDC4B.

Principal component analysis

Principal component analysis (PCA) was carried out using the `smartpca` program from EIGENSOFT (v6.0.1; https://github.com/DReichLab/EIG), on the Serbian genome combined with the SNV data (600,841 loci) from Lazaridis et al. [3]. Only the subset of European individuals from their curated fully public dataset was used, reducing the original set of 1,964

individuals to 260. A projection to the first two principal components was used to establish the correspondence between genetic and geographic distance in our results.

Neanderthal introgression

To test for Neanderthal introgression in the Serbian genome, we computed D-statistics [49, 50] using this genome and the dataset from Lazaridis et al. [9]. This dataset includes 294 ancient individuals (only one of which was used here) and a diverse set of 2,068 present-day humans, genotyped on the Affymetrix Human Origins array. Both the archaic and modern genotype data were provided in the PACKEDANCESTRYMAP format, and were combined using the `mergeit` program from EIGENSOFT (v6.1.2; https://github.com/DReichLab/EIG). The merged dataset, in total, contains 2,362 samples genotyped at 621,799 SNV loci. Upon request, we completed the consent form and obtained approval from David Reich's laboratory before using this dataset. Some individuals from the study of Lazaridis et al. [9] could not be included due to consent issues relating to data distribution.

We next genotyped the Serbian genome against these predefined SNVs using GATK HaplotypeCaller and following the GATK best practices recommendations [34, 35]. We converted the resulting VCF files to the EIGENSTRAT format using VCFtools (v0.1.12a, [51]), and integrated the Serbian genotype with the modern and ancient datasets. Finally, we ran qpDstat from AdmixTools (default setting, v701) to calculate D-statistics and to test for Neanderthal gene flow into the Serbian genome [50].

Burden of pathogenic variation

Variants of putative clinical significance were identified using genotype-phenotype databases as well as computational variant-effect prediction. Manually curated genotype-phenotype databases, such as the Human Gene Mutation Database (HGMD) [52], ClinVar [53] and PharmGKB [54], annotate variants with a known relationship to phenotype [52, 55]. Clinical Annotations from PharmGKB were compared against dbSNP v142 rsIDs [56] obtained using the annotate_variation.pl script in ANNOVAR and `avsnp142`. Variants identified by GATK were compared against HGMD and ClinVar to identify potentially disease-causing and disease-associated mutations.

All variants in protein-coding regions were extracted and inputted to the MutPred suite of tools [57–60]. The remaining variation observed in the proband was interrogated using CADD [61]. For disease and gene ontology associations, the hypergeometric test in WebGestalt was used with Benjamini-Hochberg correction for multiple hypothesis-testing [62]. The background set that was used for these analyses included all protein-coding genes from the human reference genome. For the significance of an ontology term to be confirmed, at least five genes were required to be associated with it.

Results

Effect of genotyping software

The choice of computational tools and their parameters in processing raw sequencing reads can significantly impact the resulting genome and the entirety of subsequent analysis [63, 64]. To understand the uncertainty of variant identification in our subject, we evaluated two different read mappers, BWA-MEM [29] and Bowtie2 [30], and two different variant callers, GATK [31] and Platypus [32].

The results from four different platforms are compared and contrasted in Fig 1. The SNV calling shows good concordance between both read mappers and variant callers, with a large

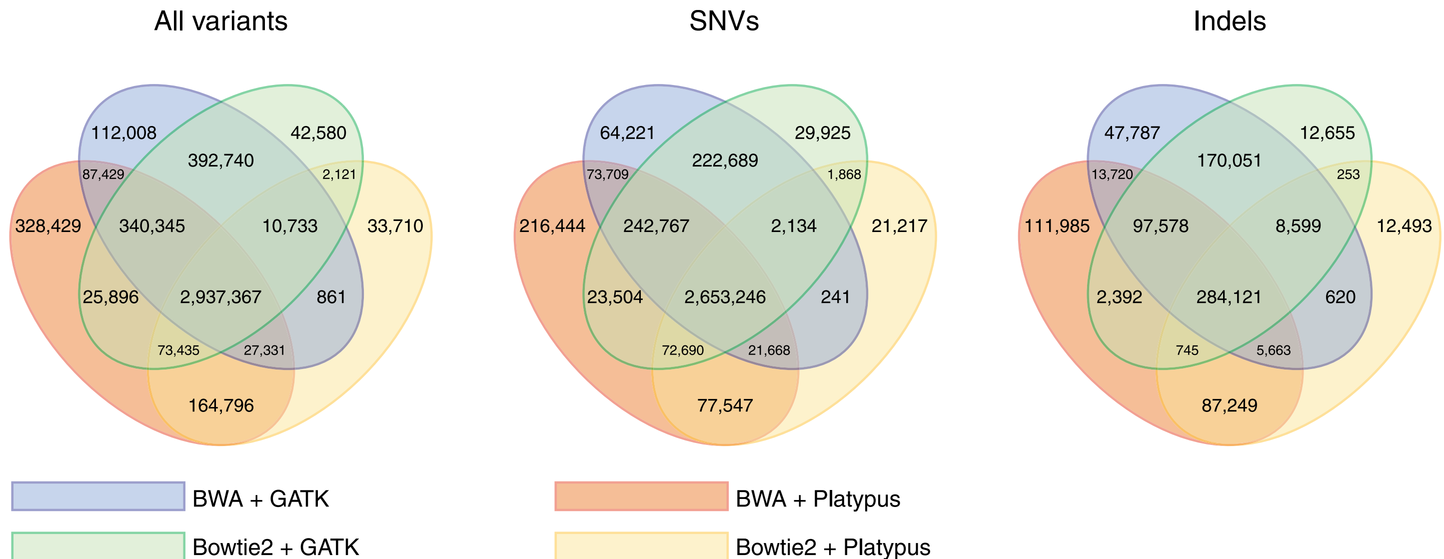


Fig 1. Venn diagrams showing the total numbers of identified variants using two read mappers (BWA [29], Bowtie2 [30]) and two variant callers (GATK [31], Platypus [32]).

https://doi.org/10.1371/journal.pone.0208901.g001

proportion of variants identified by either platform being identified by all platforms. Using the BWA-MEM mapper (which we refer to simply as "BWA" from now on), for example, 2,991,390/3,280,434 = 91.2% of SNVs identified by GATK were also identified by Platypus and 89.1% of SNVs identified by Platypus were also identified by GATK (Fig 1). Indel calling, on the other hand, is less reliable, with 401,082/627,519 = 63.9% variants identified by GATK also identified by Platypus and only 66.7% of variants identified by Platypus being also identified by GATK. The influence of read mappers was markedly lower; i.e., using the GATK variant caller, we found that 95.1% of SNVs and 89.3% of indels identified with BWA were also identified with Bowtie2, and 98.3% SNVs and of 97.6% of indels identified with Bowtie2 were also identified with BWA. Smaller percentages of overlap were observed for Platypus. Based on the results observed in this work (Table A in S1 File) and the extent of usage of these tools in resequencing human genomes, we selected BWA+ GATK as our main platform.

Identification of genetic variants

The genome of a Serbian individual was sequenced according to the protocols described in the Materials and Methods, with all 22 autosomes having similar coverage and the X and Y chromosome having approximately half this coverage. The genome sequencing and mapping achieved an average read depth of 34.7, with 98.3% of GRCh37 reference bases having coverage of 10-fold or more and 89.4% having coverage of 20-fold or more. The number of zero-depth positions were 7,649,443 (0.3%). The coverage distribution is shown in the Supporting Information (S1 Fig).

Using the BWA+ GATK pipeline, we identified a total of 3,908,814 variants (83.9% SNVs, 16.1% indels; Fig 1) in the Serbian genome, of which 2,195,638 (56.2%) were heterozygous with one non-reference allele, 23,095 (0.6%) were heterozygous with two non-reference alleles, and 1,690,081 (43.2%) were homozygous for a non-reference allele. The reported variants passed all quality filters of GATK (marked as "PASS") and were subsequently mapped to GRCh37 human reference genomic regions using ANNOVAR [36]. It is important to mention that ANNOVAR considers all heterozygous positions with both alternative alleles as two

Table 1. Summary of identified variants using BWA+ GATK. Variants not present in gnomAD [66] are listed as novel and variants identified by all four genotyping platforms are listed as confident.

Type of Variant	Variant	Novel	Confident variants	Confident novel
upstream	23094	320	16211	90
upstream; downstream	881	8	624	4
UTR5	5205	54	4055	22
UTR5; UTR3	16	1	12	0
exonic	20706	145	17114	115
exonic; splicing	33	1	22	0
splicing	151	0	107	0
intronic	1410507	20531	1078226	4336
UTR3	31066	409	24095	101
downstream	26685	398	19351	61
ncRNA_exonic	13064	129	9520	30
ncRNA_exonic; splicing	3	0	2	0
ncRNA_intronic	235936	3376	173168	832
ncRNA_splicing	65	1	51	0
ncRNA_UTR5	1	1	0	0
intergenic	2164496	34779	1597484	6848

https://doi.org/10.1371/journal.pone.0208901.t001

different variants. Mechanisms by which heterozygous alternative alleles can arise include sequencing errors and highly variable sites, some of which are tri-allelic because of rare mutational events [65]. Therefore, the resulting genome contains a total of 3,931,909 variants, of which 2,940,042 (74.8%) were identified by all four platforms and are considered to be confident identifications. Unsurprisingly, the majority of identified variants were found to reside in the more expansive and less evolutionarily constrained intergenic and intronic regions (Table 1).

To identify novel variation, we compared the identified variants against the Genome Aggregation Database (gnomAD) [66]. We found that 1.5% (60,153) all variants and 0.4% (12,439) of confident variants were not present in gnomAD. We shall refer to these variants as "novel" and "confident novel" variants, respectively. The breakdown of all variants and novel variants with respect to genomic location is shown in Tables 1 and 2. The percentage of novel variants varied across categories, comprising 0.9% (80) of nonsynonymous variants, 0.4% of synonymous variants, 0.7% (145) of exonic variants, 1.5% (20,531) of intronic variants, and 1.6%

Table 2. Summary of identified exonic variants using BWA+GATK. Variants not present in gnomAD [66] are listed as novel and variants identified by all four platforms are listed as confident.

Type of Variant	Variants	Novel	Confident variants	Confident novel
synonymous SNV	10381	42	8965	36
nonsynonymous SNV	9328	80	7559	69
nonframeshift deletion	137	2	62	0
nonframeshift insertion	117	3	58	0
frameshift deletion	103	6	45	4
frameshift insertion	74	3	37	1
stopgain	87	6	54	4
stoploss	11	0	9	0
unknown	501	4	347	1

https://doi.org/10.1371/journal.pone.0208901.t002

(34,779) of intergenic variants. We found that 45.0% (9,328/20,739) of the exonic variants were nonsynonymous, whereas 50.1% (10,381/20,739) were synonymous. Similar fractions were observed for the confident variants (44.1% vs. 52.3%). Of the 3,871,756 GATK variants that are also observed in the gnomAD database, 3,805,264 (95%) of these variants are annotated to have allele frequency greater than 1% in gnomAD and 3,676,638 (95%) with allele frequency greater than 5%. The proportion of novel variation in the Serbian individual is at the lower end of the distribution compared to 1000 Genomes Project participants (S6 Fig), consistent with a significantly larger size of gnomAD that currently integrates 15,708 whole-genomes and 125,748 exomes.

Using SVE and FusorSV, we identified 848 deletions and 3 duplications, which include the most confident calls generated by FusorSV after merging call-sets from seven different SV-callers using the default fusion model. The numbers of structural variants called by individual SV-callers are reported in (Table B in S1 File). The deletions in the Serbian genome have a length distribution (S7 Fig) similar to the deletions in the 27 deep-coverage samples of the 1000 Genomes Project reported by FusorSV [38]. The lengths of the three duplications are 313101, 362391 and 471821 bp. We used SVint to annotate the functional impact of the structural variants. The genes that overlap with the identified structural variants are listed in S1 File Tables C and D.

Genetic variation and geographic distance

The projection of the Serbian individual to the first and second principal components against European groups from [3] confirms that individuals from the same geographic region cluster together (Fig 2). We clearly distinguish clusters of major populations composed of individuals from the same region, approximately mirroring a map of Europe. The PCA plot demonstrates

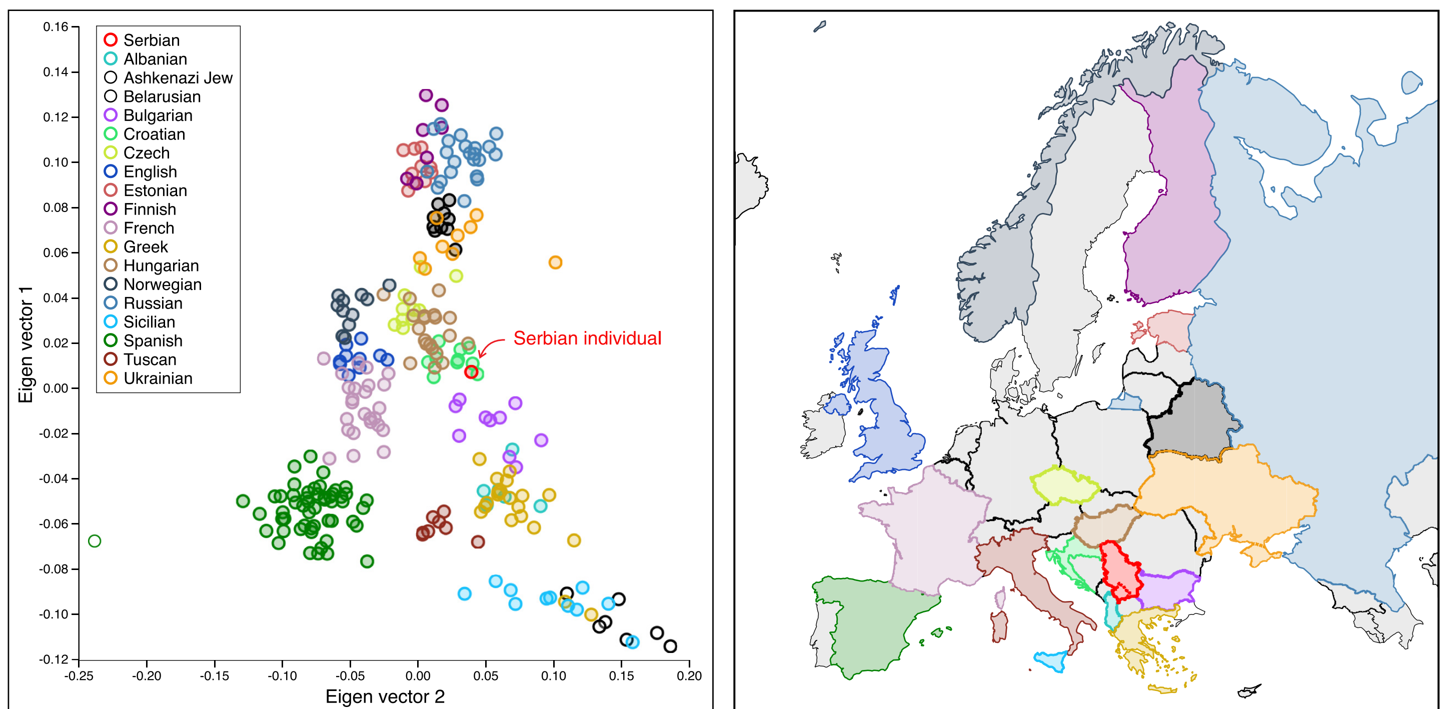


Fig 2. Principal component analysis (PCA) plot showing the proximity of the genome sequenced in this study to other European genomes. As observed in previous studies [2, 3], genomic distance correlates with geographic distance.

https://doi.org/10.1371/journal.pone.0208901.g002

that the genetic ancestry of the Serbian individual analyzed in the present study corresponds to its geographic distance from other populations. It is positioned in close proximity of the Croatian, Bulgarian, and Hungarian populations.

A somewhat surprising finding is the similarity of distances between the Serbian individual and other mostly Slavic populations (Russian, Belarus, Ukrainian) relative to distances to various Central, Western, and Southern European groups (Czech, French, English, Albanian, Greek). The average Euclidean distance and variance between the Serbian individual and each of the available populations in the two-dimensional space of major PCA components is as follows: Croatian (0.016826 ± 0.010526), Bulgarian (0.033603 ± 0.000225), Hungarian (0.037121 ± 0.000177), Czech (0.053687 ± 0.000033), Albanian (0.058875 ± 0.000117), Ukrainian (0.064328 ± 0.000062), Belarusian (0.069803 ± 0.000043), Greek (0.071108 ± 0.00062), Tuscan (0.0736441 ± 0.000028), French (0.083077 ± 0.000159), English (0.084570 ± 0.000142), Norwegian (0.092721 ± 0.00088), Russian (0.095968 ± 0.000079), Estonian (0.098421 ± 0.000046), Finnish (0.108523 ± 0.000154), Sicilian (0.120370 ± 0.000481), Spanish (0.134602 ± 0.000776), Ashkenazi (0.156692 ± 0.000538). The three closest individuals to the Serbian genome were of Croatian ancestry (0.0038, 0.0046, and 0.0108).

We note that combining the Serbian individual with the set of 260 European individuals from Lazaridis et al. [3] caused 50 formerly biallelic sites to become triallelic (no monoallelic sites became triallelic). The triallelic sites were removed from the analysis, leaving 600,791 sites in the analysis. The `smartpca` program was applied to the 261-by-600,791 genotype matrix.

Gene flow with Neanderthals

Comparisons between Neanderthals and modern humans have previously revealed evidence of gene flow from Neanderthals to Europeans [49, 50, 67, 68]. To test whether the Serbian genome shares an excess of alleles with the Neanderthal genome, we integrated the Serbian genotype with a published panel of ancient and modern humans (Materials and Methods). We calculated D-statistics as a formal test for gene flow based on a four-taxon phylogeny, $D(P_1, P_2, P_3, O)$, where P_i ($i \in \{1, 2, 3\}$) are populations and O is an outgroup. Given a scenario where gene flow is absent, the derived alleles of P_3 are expected, with equal likelihood, to match those of P_1 and P_2; i.e., $D = 0$. Alternatively, either P_1 or P_2 could share alleles with P_3 more often than not, in which case D deviates from zero.

We computed *D*(Yoruba, Serbian, Altai, Chimpanzee) for testing for gene flow between Neanderthals ("Altai") and the given Serbian genome. We expected a positive *D* value, given previous evidence that Neanderthals exchanged more alleles with Europeans than with Africans. The test returned a *D* value of 0.0241 ± 0.004476, which significantly deviated from zero (Z-score = 5.39; Table 3), suggesting gene flow between Neanderthal and the lineage leading to the Serbian genome. To validate this result, we also ran the test for other European populations (Table 3). D-statistics calculated for Croatian, French, Greek and Russian genomes were comparable to our result, all falling within the expected range of values reported in previous studies [49, 67, 68].

We further attempted to ensure that the calculated D-statistics were unbiased. To do this, we repeated the analysis by replacing Yoruba with Mbuti, as some of the Yoruba samples could have had some recent European admixture. The calculation for *D*(Mbuti, Serbian, Altai, Chimpanzee) yielded a *D* value of 0.0186 ± 0.004763 (Z-score = 3.99; Table 3), consistent with our results using the Yoruba samples. We next checked whether the Serbian individual has reference biases in genotyping that could have inflated the *D* value. We performed D-statistics tests in the form of *D*(other European population, Serbian, Mbuti, hg19ref) and chose Croatian, French, Greek and Russian as the "other European population". We obtained no test

Table 3. Testing gene flow with Neanderthals. The results show the D-statistic (*D*), its standard error (SE) and Z-score (*Z*) for the test using the set of populations P_1, P_2, and P_3, with Chimpanzee as an outgroup (*O*). The last two columns show ABBA vs. BABA counts over the four genomes (P_1, P_2, P_3, *O*).

P_1	P_2	P_3	*O*	*D*	SE	Z-score	ABBA	BABA
Yoruba	Serbian	Altai	Chimpanzee	0.0241	0.004476	5.393	18158	17302
Yoruba	Croatian	Altai	Chimpanzee	0.0233	0.003192	7.302	18268	17436
Yoruba	French	Altai	Chimpanzee	0.0266	0.003012	8.821	18284	17338
Yoruba	Greek	Altai	Chimpanzee	0.0270	0.003034	8.906	18266	17305
Yoruba	Russian	Altai	Chimpanzee	0.0288	0.003096	9.306	18328	17302
Mbuti	Serbian	Altai	Chimpanzee	0.0186	0.004763	3.909	18817	18129
Mbuti	Croatian	Altai	Chimpanzee	0.0178	0.003693	4.832	18891	18229
Mbuti	French	Altai	Chimpanzee	0.0210	0.003532	5.941	18902	18125
Mbuti	Greek	Altai	Chimpanzee	0.0214	0.003578	5.978	18897	18106
Mbuti	Russian	Altai	Chimpanzee	0.0232	0.003600	6.434	18932	18074

https://doi.org/10.1371/journal.pone.0208901.t003

results indicating the bias of Serbian genotypes toward the reference (Croatian: 0.0054 ± 0.004183; French: 0.0038 ± 0.004078; Greek: 0.0090 ± 0.004182; Russian: 0.0074 ± 0.004192).

Analysis of medically relevant variants

The sequenced genome contains 2,343 genetic variants that are present in HGMD by virtue of their having been previously associated with a risk of disease; the proportions of variants within each effect category are shown in Table 4. Several homozygous variants, manually annotated as disease-causing (DM) are observed in the genome, shown in Table 5. Of these,

Table 4. Amount of disease-causing and potentially disease-relevant variation in the Serbian genome. Identified variants were searched against HGMD and broken down into the phenotypic categories of HGMD. Variants were broken down into exonic and noncoding as well as homozygous and heterozygous.

	Exome		Noncoding	
	Hom	Het	Hom	Het
Disease-causing mutations (DM)	1	9	4	6
Likely disease-causing mutations (DM?)	29	51	8	31
Disease-associated polymorphisms with additional supporting functional evidence (DFP)	78	139	203	301
Disease-associated polymorphisms (DP)	233	356	189	322
Polymorphisms that affect gene/protein structure, function or expression but with no reported disease association (FP)	63	95	95	130

The number of homozygous and heterozygous variants that are associated with variants reported in HGMD. HGMD labels correspond to the strength and/or evidence for the relationship between variant and disease.

https://doi.org/10.1371/journal.pone.0208901.t004

Table 5. Disease-causing variants observed in the proband. The table summarizes the analysis of five homozygous variants form the sequenced genome that are listed by HGMD as disease-causing.

Gene	Variant	rsID	Phenotype
MIR137HG	NC_000001.10:g.98502934G>T	rs1625579	Schizophrenia increased risk
SLC12A3	NM_000339.2:c.1670-8C>T	NA	Gitelman syndrome without hypomagnesaemia
DUOXA2	NM_207581.3:c.554+6C>T	NA	Hypothyroidism
F13A1	NM_000129.3:c.-19+12C>A	rs2815822	Factor XIII deficiency
PNPLA2	NP_065109.1:p.P481L	rs1138693	Myopathy late-onset

https://doi.org/10.1371/journal.pone.0208901.t005

one is a youth-onset phenotype, Factor XIII deficiency, associated with homozygosity for the disease-causing allele (NM_000129.3:c.-19+12C>A) in the proband's genome. The disease phenotypes associated with these homozygous mutations typically become apparent in childhood, and therefore their occurrence in a healthy adult is indicative of variable penetrance. The other homozygous disease-causing variants result in phenotypes that have not yet been observed in either the individual or in their family history; perhaps reflecting either low expressivity or late-onset. Observed heterozygous disease-causing mutations are primarily childhood-onset without presentation in the individual, although they may represent recessive conditions; thus, their failure to manifest may not necessarily be indicative of poor reporting or curation quality. Next, we identified several variants with pathogenic annotation in the ClinVar database, an open-access alternative to HGMD [53]. These variants are either low-confidence or without known family history; more details are available in the Supporting Information (S1 File).

We also identified several variants of potential pharmacogenetic relevance using PharmGKB. Variants in PharmGKB are assigned Clinical Annotation Levels of Evidence from variants with preliminary evidence (Level 4) to high confidence variant-drug combinations with medically endorsed integration into health systems (Level A1). The genome contains a single variant with a high-confidence annotation (Level 1B): rs2228001, associated with toxicity and adverse drug reaction to cisplatin, a chemotherapeutic agent. A further 17 variants were annotated with moderate evidence to impact the dosage, efficacy, metabolism and/or toxicity of drugs for diverse phenotypes including chronic hepatitis C, organ transplantation rejection, glaucoma, depression, schizophrenia, asthma, epilepsy and HIV infections, as well as several chemotherapy drugs.

Pathogenicity prediction. In addition to known disease-associated variants, we identified missense variants predicted to be pathogenic by MutPred2 [57]. Of the 11,206 missense variants called by GATK, 9,329 passed all quality filters (annotated as 'PASS'). Of these, 9,305 variants were unambiguously mapped to the correct protein isoforms and hence were amenable for prediction by MutPred2. Based on a score threshold of 0.8 (estimated 5% false positive rate), 95 missense variants were predicted to be 'pathogenic.'

Of these, 14 variants were found in the homozygous state and 81 were found in the heterozygous state. Genes for these variants were enriched in GO terms related to peptidase activity S8 Fig). A similar analysis for disease associations revealed that the subject may be at risk for cardiovascular disorders (Table I in S1 File).

Next, we applied computational predictors on the remaining protein coding variation with the MutPred family of tools. First, we assessed the pathogenicity of 180 nonsense and frameshifting insertion and deletion variants with MutPred-LOF [58]. From this set, we identified a total of 7 variants with scores above the 0.5 score threshold (corresponding to a 5% false positive rate) (Table E in S1 File). Next, we assessed 279 non-frameshifting insertion and deletion variants with MutPred-Indel and identified 12 variants described in (Table F in S1 File. Finally, we assessed the pathogenicity of the 90 SNV splicing variants with MutPred Splice [59]. Of these, 28 of the variants scored at least 0.6 and were therefore classified as a "Splice Affecting Variant" by MutPred Splice. One of these variants is predicted to cause loss of natural 3' splice sites, two variants are predicted to interrupt cryptic 3' splice sites, and three variants are predicted to disrupt cryptic 5' splice sites, described in the Supporting Information (Table G in S1 File).

To ensure assessment of the complete variome of the proband, we utilized CADD v1.3 [61] to evaluate all noncoding variants. To do this, we utilized a scaled C-score cutoff of 20 to identify the 1% most damaging variants. In total, we found 16 UTR variants, 1,630 intronic variants, 3,911 intergenic variants, 80 regulatory variants, 839/533 upstream/downstream variants,

and 9 variants annotated as "noncoding_change." All of these were predicted to be deleterious. The noncoding variants with the highest C-scores are described in the Supporting Information (Table H in S1 File).

Discussion

This work describes the first whole-genome sequencing of a Serbian individual. Ancestry analysis positioned the Serbian individual in closest proximity to the Croatian population, consistent with its Southern Slavic ancestry [69]. Our analyses further support the hypothesis of gene flow between Neanderthal and pan-European ancestral populations, with the level of introgression into the Serbian genome being within the range observed in other European populations. Previous genetic studies involving Slavic populations employed mitochondrial, Y-chromosome and SNV-panel data to investigate the relationship between geographic, genetic and linguistic distances [69, 70]. Consistent with this work, our analyses expand the scope beyond Slavic populations and further contribute to the understanding of human genetic variation and its geographic distribution.

In contrast to studies using genotyping arrays [2, 3, 69, 70], the availability of whole-genome sequences presents the opportunity for a high-resolution individualized analysis. To this end, we found that the sequenced genome contains a significant number of previously unobserved variants, which emphasizes the importance of continued sequencing of a large number of individuals, especially from previously uncharacterized ethnic groups. Subsequent sequencing of other Serbian individuals could provide further insight into these novel variants; e.g., whether they are private to the population or to the individual. Such results would in turn contribute important information regarding variants that are currently considered to be rare, with implications for improved variant interpretation. Furthermore, new algorithms and reduced sequencing costs will have the potential to provide higher-quality analysis of structural variants. Our analysis also found a number of variants of clinical and pharmacogenomic significance that might extend beyond an individual's disease risks to facilitate possible future medical interventions although conclusions are limited without validation and knowledge of allele frequencies in the Serbian population [71, 72]. Such variants might contribute to better outcomes in studies of disease penetrance, mechanistic understanding of population risks, and database curation.

Recent advances in high-throughput sequencing and reduced costs of genotyping have greatly facilitated whole-genome data generation, and have become key to understanding both human phenotypes and early human history [2, 3]. However, modern technology and cost structure continue to pose challenges in determining and interpreting one's genome [73]. Variation in read mapping and variant calling contribute to the uncertainty of interpretation with different software packages, identifying different sets of variants. We found that inter-software discrepancies ranged from relatively small for SNVs to considerable for insertions and deletions, especially for structural variants. Therefore, variant and genome interpretation demand caution, since thousands of SNVs and tens of thousands of indels may simply constitute genotyping errors [74, 75].

It is worth mentioning that in addition to the technical aspects of genome sequencing, an important aspect of genome interpretation concerns psychosocial uncertainty due to phenotypic and privacy-associated risks [76]. The geographic distance analysis in this study has provided evidence that supports the individual's own sense of Serbian ancestry; however, the finding of multiple predicted youth-onset pathogenic mutations in a healthy individual provides cautionary lessons for predictive medicine.

Supporting information

S1 Fig. Histogram of read depths.
(EPS)

S2 Fig. Read depth across SNV and insertion/deletion variants across the four pipelines.
(EPS)

S3 Fig. Size of insertion/deletion variants across the four pipelines.
(EPS)

S4 Fig. Number of heterozygous and homozygous variants across the four pipelines.
(EPS)

S5 Fig. Size of insertion/deletion variants in the Serbian genome.
(EPS)

S6 Fig. Proportion of novel variants in thousand genomes project participants.
(EPS)

S7 Fig. Length distribution of deletions called by FusorSV.
(EPS)

S8 Fig. GO terms enriched in the set of 81 genes that harbored the 95 missense variants predicted to be pathogenic.
(EPS)

S1 File. Annotation descriptions and tables of variants scored as pathogenic.
(PDF)

Acknowledgments

We thank Dr. Chuanchao Wang for his suggestions and discussion on testing for Neanderthal admixture in the Serbian genome. We also thank the Associate Editor and anonymous reviewers for their insightful comments that have improved the precision and quality of the paper. This work was supported by the Center for Bioinformatics Research at Indiana University, and was originally carried out as part of a class project for INFO-I590.

Author Contributions

Methodology: Kymberleigh A. Pagel, Vikas Pejaver, Simo V. Zhang.

Supervision: Predrag Radivojac.

Writing – original draft: Wazim Mohammed Ismail, Kymberleigh A. Pagel, Vikas Pejaver, Simo V. Zhang, Sofia Casasa, Predrag Radivojac.

Writing – review & editing: Wazim Mohammed Ismail, Kymberleigh A. Pagel, Vikas Pejaver, Simo V. Zhang, Matthew Mort, David N. Cooper, Matthew W. Hahn, Predrag Radivojac.

References

1. Consortium GP, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. https://doi.org/10.1038/nature15393
2. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008; 456(7218):98–101. https://doi.org/10.1038/nature07331 PMID: 18758442

3. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014; 513(7518):409–413. https://doi.org/10.1038/nature13673 PMID: 25230663
4. Pagani L, Lawson DJ, Jagoda E, Morseburg A, Eriksson A, Mitt M, et al. Genomic analyses inform on migration events during the peopling of Eurasia. Nature. 2016; 538(7624):238–242. https://doi.org/10.1038/nature19792 PMID: 27654910
5. Montinaro F, Busby GB, Gonzalez-Santos M, Oosthuitzen O, Oosthuitzen E, Anagnostou P, et al. Complex ancient genetic structure and cultural transitions in southern African populations. Genetics. 2017; 205(1):303–316. https://doi.org/10.1534/genetics.116.189209 PMID: 27838627
6. House GL, Hahn MW. Evaluating methods to visualize patterns of genetic differentiation on a landscape. Mol Ecol Resour. 2018; 18(3):448–460. https://doi.org/10.1111/1755-0998.12747 PMID: 29282875
7. Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, et al. The importance of race and ethnic background in biomedical research and clinical practice. N Engl J Med. 2003; 348(12):1170–1175. https://doi.org/10.1056/NEJMsb025007 PMID: 12646676
8. Gibson G, Muse SV. A primer of genome science. 3rd ed. Sunderland, Mass.: Sinauer Associates; 2009.
9. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, et al. Genomic insights into the origin of farming in the ancient Near East. Nature. 2016; 536(7617):419–424. https://doi.org/10.1038/nature19310 PMID: 27459054
10. Manrai AK, Funke BH, Rehm HL, Olesen MS, Maron BA, Szolovits P, et al. Genetic misdiagnoses and the potential for health disparities. N Engl J Med. 2016; 375(7):655–665. https://doi.org/10.1056/NEJMsa1507092 PMID: 27532831
11. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007; 5(10):e254. https://doi.org/10.1371/journal.pbio.0050254 PMID: 17803354
12. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008; 452(7189):872–876. https://doi.org/10.1038/nature06884 PMID: 18421352
13. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, et al. The diploid genome sequence of an Asian individual. Nature. 2008; 456(7218):60–65. https://doi.org/10.1038/nature07484 PMID: 18987735
14. Chekanov NN, Boulygina ES, Beletskiy AV, Prokhortchouk EB, Skryabin KG. Individual genome of the Russian male: SNP calling and a *de novo* assembly of unmapped reads. Acta Naturae. 2010; 2(3):122–126. PMID: 22649659
15. Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, et al. Complete Khoisan and Bantu genomes from southern Africa. Nature. 2010; 463(7283):943–947. https://doi.org/10.1038/nature08795 PMID: 20164927
16. Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, et al. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nat Genet. 2010; 42(11):931–936. https://doi.org/10.1038/ng.691 PMID: 20972442
17. Suk EK, McEwen GK, Duitama J, Nowick K, Schulz S, Palczewski S, et al. A comprehensively molecular haplotype-resolved genome of a European individual. Genome Res. 2011; 21(10):1672–1685. https://doi.org/10.1101/gr.125047.111 PMID: 21813624
18. Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, et al. Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat Biotechnol. 2011; 29(1):59–63. https://doi.org/10.1038/nbt.1740 PMID: 21170042
19. Lilleoja R, Sarapik A, Reimann E, Reemann P, Jaakma U, Vasar E, et al. Sequencing and annotated analysis of an Estonian human genome. Gene. 2012; 493(1):69–76. https://doi.org/10.1016/j.gene.2011.11.022 PMID: 22138481
20. Azim MK, Yang C, Yan Z, Choudhary MI, Khan A, Sun X, et al. Complete genome sequencing and variant analysis of a Pakistani individual. J Hum Genet. 2013; 58(9):622–626. https://doi.org/10.1038/jhg.2013.72 PMID: 23842039
21. Bai H, Guo X, Zhang D, Narisu N, Bu J, Jirimutu J, et al. The genome of a Mongolian individual reveals the genetic imprints of Mongolians on modern human populations. Genome Biol Evol. 2014; 6(12):3122–3136. https://doi.org/10.1093/gbe/evu242 PMID: 25377941
22. Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N, et al. The promise of discovering population-specific disease-associated genes in South Asia. Nat Genet. 2017; 49(9):1403–1407. https://doi.org/10.1038/ng.3917 PMID: 28714977

23. Smyth N, Ramsay M, Raal FJ. Population specific genetic heterogeneity of familial hypercholesterolemia in South Africa. Curr Opin Lipidol. 2018; 29(2):72–79. https://doi.org/10.1097/MOL.0000000000000488 PMID: 29369830
24. Lencz T, Yu J, Palmer C, Carmi S, Ben-Avraham D, Barzilai N, et al. High-depth whole genome sequencing of an Ashkenazi Jewish reference panel: enhancing sensitivity, accuracy, and imputation. Hum Genet. 2018; 137(4):343–355. https://doi.org/10.1007/s00439-018-1886-z PMID: 29705978
25. Guda K, Veigl ML, Varadan V, Nosrati A, Ravi L, Lutterbaugh J, et al. Novel recurrently mutated genes in African American colon cancers. Proc Natl Acad Sci U S A. 2015; 112(4):1149–1154. https://doi.org/10.1073/pnas.1417064112 PMID: 25583493
26. Cho YS, Kim H, Kim HM, Jho S, Jun J, Lee YJ, et al. An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. Nat Commun. 2016; 7:13637. https://doi.org/10.1038/ncomms13637 PMID: 27882922
27. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016; 538(7624):161–164. https://doi.org/10.1038/538161a PMID: 27734877
28. Ball MP, Thakuria JV, Zaranek AW, Clegg T, Rosenbaum AM, Wu X, et al. A public resource facilitating clinical use of genomes. Proc Natl Acad Sci U S A. 2012; 109(30):11920–11927. https://doi.org/10.1073/pnas.1201904109 PMID: 22797899
29. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168
30. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9(4):357–359. https://doi.org/10.1038/nmeth.1923 PMID: 22388286
31. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–1303. https://doi.org/10.1101/gr.107524.110 PMID: 20644199
32. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Consortium W, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014; 46(8):912–918. https://doi.org/10.1038/ng.3036 PMID: 25017105
33. Picard Tools;. http://broadinstitute.github.io/picard/.
34. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013; 43:11.10.1–11.10.33.
35. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43(5):491–498. https://doi.org/10.1038/ng.806 PMID: 21478889
36. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38(16):e164. https://doi.org/10.1093/nar/gkq603 PMID: 20601685
37. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. https://doi.org/10.1038/nature15393 PMID: 26432245
38. Becker T, Lee WP, Leone J, Zhu Q, Zhang C, Liu S, et al. FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. Genome Biol. 2018; 19(1):38. https://doi.org/10.1186/s13059-018-1404-6 PMID: 29559002
39. Fan X, Abbott TE, Larson D, Chen K. BreakDancer: identification of genomic structural variation from paired-end read mapping. Curr Protoc Bioinformatics. 2014; 45(1):15.6.1–15.6.11.
40. Lam HY, Mu XJ, Stutz AM, Tanzer A, Cayting PD, Snyder M. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol. 2010; 28(1):47–55. https://doi.org/10.1038/nbt.1600 PMID: 20037582
41. Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 2012; 40(9):e69. https://doi.org/10.1093/nar/gks003 PMID: 22302147
42. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011; 21(6):974–984. https://doi.org/10.1101/gr.114876.110 PMID: 21324876
43. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012; 28(18):i333–i339. https://doi.org/10.1093/bioinformatics/bts378 PMID: 22962449

44. Handsaker RE, Korn JM, Nemesh J, McCarroll SA. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet. 2011; 43(3):269–276. https://doi.org/10.1038/ng.768 PMID: 21317889
45. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM. Large multiallelic copy number variations in humans. Nat Genet. 2015; 47(3):296–303. https://doi.org/10.1038/ng.3200 PMID: 25621458
46. Lindberg MR, Hall IM, Quinlan AR. Population-based structural variation discovery with Hydra-Multi. Bioinformatics. 2015; 31(8):1286–1289. https://doi.org/10.1093/bioinformatics/btu771 PMID: 25527832
47. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014; 15(6):R84. https://doi.org/10.1186/gb-2014-15-6-r84 PMID: 24970577
48. SVint, a light-weight tool for annotating structure variants located outside the coding genome;. http://compbio.berkeley.edu/proj/svint/.
49. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010; 328(5979):710–722. https://doi.org/10.1126/science.1188021
50. Slatkin M, Racimo F. Ancient DNA and human history. Proc Natl Acad Sci U S A. 2016; 113(23):6380–6387. https://doi.org/10.1073/pnas.1524306113 PMID: 27274045
51. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27(15):2156–2158. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522
52. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 2017; 136(6):665–677. https://doi.org/10.1007/s00439-017-1779-6 PMID: 28349240
53. Landrum MJ, Lee JM, M B, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016; 44(D1):D862–D868. https://doi.org/10.1093/nar/gkv1222 PMID: 26582918
54. Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, Hewett M, et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. Pharmacogenomics J. 2001; 1(3):167–170. PMID: 11908751
55. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012; 92(4):414–417. https://doi.org/10.1038/clpt.2012.96 PMID: 22992668
56. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29(1):308–311. https://doi.org/10.1093/nar/29.1.308 PMID: 11125122
57. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, et al. MutPred2: inferring the molecular and phenotypic impact of amino acid variants. bioRxiv 134981. 2017;.
58. Pagel KA, Pejaver V, Lin GN, Nam HJ, Mort M, Cooper DN, et al. When loss-of-function is loss of function: assessing mutational signatures and impact of loss-of-function genetic variants. Bioinformatics. 2017; 33(14):i389–i398. https://doi.org/10.1093/bioinformatics/btx272 PMID: 28882004
59. Mort M, Sterne-Weiler T, Li B, Ball EV, Cooper DN, Radivojac P, et al. MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. Genome Biol. 2014; 15(1):R19. https://doi.org/10.1186/gb-2014-15-1-r19 PMID: 24451234
60. Pagel KA, Mort M, Cooper DN, Mooney SD, Radivojac P. Pathogenicity and functional effects of non-frameshifting insertion/deletion variation in the human genome. Unpublished;.
61. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46(3):310–315. https://doi.org/10.1038/ng.2892 PMID: 24487276
62. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res. 2013; 41(Web Server issue):77–83. https://doi.org/10.1093/nar/gkt439
63. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015; 5:17875. https://doi.org/10.1038/srep17875 PMID: 26639839
64. Zook J, McDaniel J, Parikh H, Heaton H, Irvine SA, Trigg L, et al. Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials. bioRxiv 281006. 2018;.
65. Hodgkinson A, Eyre-Walker A. Human triallelic sites: evidence for a new mutational mechanism? Genetics. 2010; 184(1):233–241. https://doi.org/10.1534/genetics.109.110510 PMID: 19884308

66. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536(7616):285–291. https://doi.org/10.1038/nature19057 PMID: 27535533
67. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014; 505(7481):43–49. https://doi.org/10.1038/nature12886 PMID: 24352235
68. Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. Mol Biol Evol. 2011; 28(8):2239–2252. https://doi.org/10.1093/molbev/msr048 PMID: 21325092
69. Kushniarevich A, Utevska O, Chuhryaeva M, Agdzhoyan A, Dibirova K, Uktveryte I, et al. Genetic heritage of the Balto-Slavic speaking populations: a synthesis of autosomal, mitochondrial and Y-chromosomal data. PLoS One. 2015; 10(9):e0135820. https://doi.org/10.1371/journal.pone.0135820 PMID: 26332464
70. Davidovic S, Malyarchuk B, Aleksic J, Derenko M, Topalovic V, Litvinov A, et al. Mitochondrial super-haplogroup U diversity in Serbians. Ann Hum Biol. 2017; 44(5):408–418. https://doi.org/10.1080/03014460.2017.1287954 PMID: 28140657
71. Ramos E, Doumatey A, Elkahloun AG, Shriner D, Huang H, Chen G, et al. Pharmacogenomics, ancestry and clinical decision making for global populations. Pharmacogenomics J. 2014; 14(3):217–222. https://doi.org/10.1038/tpj.2013.24 PMID: 23835662
72. Wright GEB, Carleton B, Hayden MR, Ross CJD. The global spectrum of protein-coding pharmacogenomic diversity. Pharmacogenomics J. 2018; 18(1):187–195. https://doi.org/10.1038/tpj.2016.77 PMID: 27779249
73. van Nimwegen KJ, van Soest RA, Veltman JA, Nelen MR, van der Wilt GJ, Vissers LE, et al. Is the $1000 genome as near as we think? A cost analysis of next-generation sequencing. Clin Chem. 2016; 62(11):1458–1464. https://doi.org/10.1373/clinchem.2016.258632 PMID: 27630156
74. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011; 12(6):443–451. https://doi.org/10.1038/nrg2986 PMID: 21587300
75. Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok PY, Schaefer C, et al. Estimating genotype error rates from high-coverage next-generation sequence data. Genome Res. 2014; 24(11):1734–1739. https://doi.org/10.1101/gr.168393.113 PMID: 25304867
76. Wang S, Jiang X, Singh S, Marmor R, Bonomi L, Fox D, et al. Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States. Ann N Y Acad Sci. 2017; 1387(1):73–83. https://doi.org/10.1111/nyas.13259 PMID: 27681358