Experience in developing the human genome standard E701

Iuliia Vasiliadis[1,*], Vera Belova[1], Anna Shmitko[1], Anna Kuznetsova[1], Alina Samitova[1], Oleg Suchalko[1], Andrey Goltsov[2], Peter Shatalov[3], Peter Shegai[3], Olga Melkova[4], Tatiana Kulyabina[4], Elena Kulyabina[4], Denis Rebrikov[1], Dmitriy Korostin[1]

[1] Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Pirogov Russian National Research Medical University, Ostrovityanova str. 1, Moscow, Russian Federation, 117513

[2] National Medical Research Center for Obstetrics, Gynecology and Perinatology named after Academician V.I.Kulakov, Oparina str. 4, Moscow, Russian Federation, 117997

[3] National Medical Research Radiological Centre of the Ministry of Health of the Russian Federation, Koroleva str. 4, Obninsk, Russian Federation, 249036

[4] All-Russian Research Institute for Metrological Service (VNIIMS), Ozernaya str. 46, Moscow, Russian Federation, 119361

[*] To whom correspondence should be addressed. Email: julia.vasiliadis@gmail.com

Abstract

The first Russian human genome standard E701 was developed through a collaborative research involving four laboratories: Pirogov Russian National Research Medical University, National Medical Research Center for Obstetrics, Gynecology and Perinatology named after Academician V.I.Kulakov, National Research Center Kurchatov Institute, and National Medical Research Radiological Centre. Whole-genome sequencing of short reads on various platforms (MGI Tech, Illumina) and alignment to the reference human genome GRCh38.p14 were performed for the E701 sample. Subsequently, 3842877 high confidence genomic variants were identified, which can be used as a standard for calculating statistical quality metrics while analyzing sequencing data. Furthermore, 9096 biallelic variants were identified on the autosomes and the X chromosome, with a minor allele frequency exceeding 0.4. Additionally, mitochondrial DNA sequencing was performed with the breadth of coverage over 99.9% at 1000X.

Introduction

The rapid development of next-generation sequencing (NGS) technologies imposes new demands on the reliability of identified genetic variants and the potential for errors in bioinformatic data analysis. Research shows that results obtained from different sequencing platforms and bioinformatics algorithms can vary significantly [1,2,3,4,5]. A highly accurate reference is essential for evaluating the accuracy and reproducibility of various sequencing methods. The most well-known human reference genomes

were developed by the Genome in a Bottle (GIAB) Consortium and are widely used for sequencing validation [6,7].
Using high-throughput sequencing from multiple platforms and bioinformatic analysis of the obtained data, we produced the first Russian human genome standard E701.

Materials and methods

Ethics Statement

This study conformed to the principles of the Declaration of Helsinki. The appropriate institutional review board approval for this study was obtained from the Ethics Committee at the Pirogov Russian National Research Medical University. Patient provided written informed consent for sample collection, subsequent analysis, and publication thereof.

DNA source

The genomic DNA derived from peripheral blood mononuclear cells came from a white Caucasian male with karyotype 46XY with no history of hereditary pathologies. Whole blood was collected in EDTA-Vacutainer tubes and the DNA extraction was performed using the QIAamp DNA Blood Kit (Qiagen) according to the manufacturer's protocol.

Sequencing

Whole-genome sequencing (WGS) data in paired-end (PE) mode with 30X coverage were obtained using various high-throughput sequencing platforms in four laboratories: Pirogov Russian National Research Medical University (RSMU), National Medical Research Center for Obstetrics, Gynecology and Perinatology named after Academician V.I.Kulakov (NCAGP), National Research Center Kurchatov Institute (NRCKI), and National Medical Research Radiological Centre (NMICR) (Table 1). Library preparation followed the manufacturer’s instructions (MGI and Illumina).

Laboratory	Platform	Model	Sequencing mode	Number of reads, M
Pirogov Russian National Research Medical University	MGI Tech	G-400	PE100	1416
National Medical Research Center for Obstetrics, Gynecology and Perinatology named after Academician V.I.Kulakov	Illumina	NovaSeq	PE151	535

National Research Center Kurchatov Institute	MGI Tech	G-400	PE150	947
National Medical Research Radiological Centre	MGI Tech	G-400	PE100	907

Table 1. Description of sequencing platforms and raw data quantity

Bioinformatics processing

The quality of the obtained paired FastQ files was analyzed using FastQC v0.11.9 [8]. Based on the quality metrics, the FastQ files were trimmed using BBDuk by BBMap v38.96 [9]. Reads were aligned to the indexed reference genome GRCh38.p14 (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/) using bwa-mem2 v2.2.1 [10]. SAM files were converted into BAM files and sorted using SAMtools v1.9 [11]. The number of duplicates was calculated using Picard MarkDuplicates v2.22.4 [12].

Results

Efficiency of whole-genome sequencing

To assess the alignment quality and uniformity of genome coverage, Picard CollectWgsMetrics and SAMtools flagstat were used. The results showed that over 99% of reads were aligned to the reference genome, indicating the high quality of the raw data and the efficiency of the alignment algorithm used. Median coverage of the sample obtained from each laboratory was at least 20 reads. Mean coverage for WGS E701 was 39X in the RSMU sample, 32X in the NRCKI sample, 24X in the NMICR sample, and 19X in the NCAGP sample. For all four samples, the breadth of coverage at 1X was approximately 91% and 5X about 90%. The breadth of coverage at 10X varied slightly across samples with a maximum value of 89.96% observed in the RSMU sample and a minimum value of 86.94% observed in the NCAGP sample. The values for the other metrics are presented in Supplementary table 1.

Variant calling

Variant calling was performed using two programs: bcftools mpileup v1.9 [13] and DeepVariant v1.5.0 [14]. Subsequently, multiallelic variants in VCF files were decomposed into biallelic variants using vt decompose v0.5772 [15] and then normalized using vt normalize v0.5772. A depth coverage filter (DP>=3) was applied to variants called by both programs. An additional filter (FILTER=PASS) was applied to variants called by DeepVariant, after which VCF files were merged using bcftools-1.9 merge.

Combined VCF

To identify reliable variants in the E701 sample, we filtered single nucleotide variants (SNVs) in each of the four VCF files using bcftools-1.9 view. Subsequently, the filtered VCF files were intersected using the bcftools-1.9 isec and custom-written scripts. The intersection yielded a list of 3842877 variants (Fig. 1), which served as a true set of SNVs, forming a benchmark variant call set. This approach significantly reduced the probability of including erroneous variants in the research results and ensured high reliability of the data obtained.

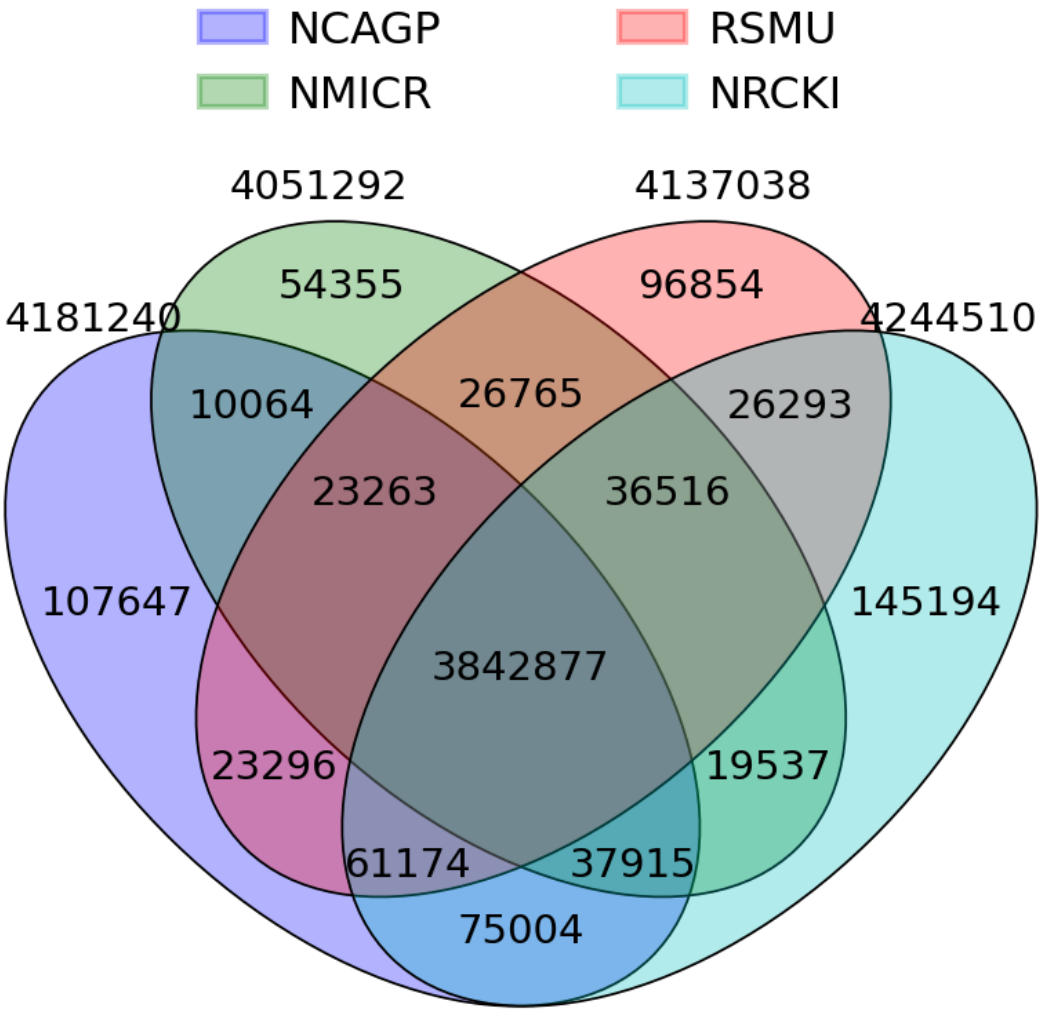

Figure 1. Venn diagram showing the overlap of variants in the E701 sample from four laboratories

Assessment of variant identification accuracy

To evaluate the results of variant calling, we used the dbSNP (Database of Single Nucleotide Polymorphisms) database release 156 (https://ftp.ncbi.nlm.nih.gov/snp/archive/b156/VCF/GCF_000001405.40.gz) [16]. The multiallelic variants from the database were split into biallelic variants and then crossed

with the variants called in the E701 sample. According to the results, 99.59% (3827006) of the variants identified in the E701 sample were present in dbSNP database (Fig. 2).

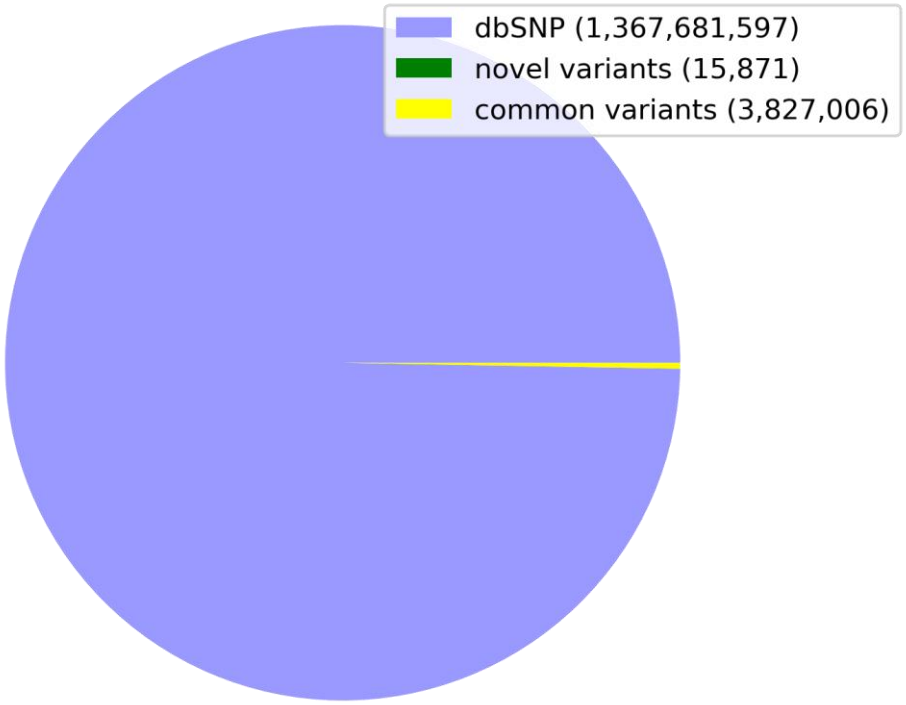

Figure 2. The number of detected SNPs in the E701 sample that overlap with variants in the dbSNP database

"EXOME-GID"

Based on the data extracted from the Genome Aggregation Database (gnomAD), a list of 9096 biallelic variants was obtained. These variants are located on autosomes and the X chromosome, within coding regions of the genome, are not associated with pathogenic traits, and have a global minor allele frequency (GMAF) in the range of 0.4 to 0.5 [17].

Ti/Tv

One of the metrics for assessing the accuracy of variant calling is the identification of point mutation types: transitions (Ti) and transversions (Tv), the ratio of which for the whole human genome is typically within the range of 2 to 2.1. The Ti/Tv ratio was calculated using the bcftools stats tool for the set of variants shared by four laboratories, yielding a value of 2.01. This result is consistent with the expected result for whole genome data.

Mitochondrial DNA

Human mitochondrial DNA (mtDNA) is a small circular molecule of about 16.6 Kb (16,569 bp) that is maternally inherited and exhibits a higher mutation rate compared to nuclear DNA [18].

For the E701-MT sample, the breadth of coverage at 1000X (PCT_1000X) was 99.994%. The mean coverage depth was 22880 reads, with a minimum of 841 reads observed at position 3107, labeled N in the mitochondrial chromosome reference assembly. The values for other metrics are presented in Supplementary table 1.

Extraction of reads unambiguously mapped to the mitochondrial chromosome and obtaining a consensus sequence based on them was performed using SAMtools-1.9. Variant calling and annotation were conducted using mutserve [19]. The variants that differ between E701-MT sample and the mtDNA reference assembly (hg38) are listed in Supplementary table 2. Variants flagged by the "artifact_prone_site" filter were excluded from the final variant list.

Conclusion

We present the first Russian human genome standard E701, which was developed through a collaborative effort involving four research laboratories (Pirogov Russian National Research Medical University, National Medical Research Center for Obstetrics, Gynecology and Perinatology named after Academician V.I.Kulakov, National Research Center Kurchatov Institute, and National Medical Research Radiological Centre). Whole-genome sequencing on various platforms was performed for the E701 sample, with sequencing and alignment efficiency evaluated against the human reference genome GRCh38.p14. We identified 3842877 genomic SNVs matched across the four laboratories, which can serve as a benchmark set to filter false positive and false negative variants generated during sequencing data processing. Of the variants identified, more than 99% were found in the dbSNP database. We also obtained a list of 9096 biallelic variants located in autosomes and the X chromosome, with a minor allele frequency exceeding 0.4. In addition to genomic DNA, mitochondrial DNA was sequenced with a breadth of coverage over 99.9% at 1000X and mean coverage depth of 22880X.

Future work will involve sequencing the family members of the E701 genetic material donor and integrating long-read sequencing technologies to improve the detection of indels and structural variants.

Data availability

Genome sequences and alignments for the E701 WGS sample in fastq.gz and BAM formats, and sequences for E701-MT mitochondrial DNA in fastq.gz have been deposited into the NCBI open-access sequence read archive (SRA) under the BioProject number PRJNA1083205.

Funding

This work was supported by Grant №075-15-2019-1789 from the Ministry of Science and Higher Education of the Russian Federation allocated to the Center for Precision Genome Editing and Genetic Technologies for Biomedicine.

REFERENCES

1. Lam, H., Clark, M., Chen, R. et al. Performance comparison of whole-genome sequencing platforms. Nat Biotechnol 30, 78–82 (2012). https://doi.org/10.1038/nbt.2065.
2. Chen, J., Li, X., Zhong, H. et al. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. Sci Rep 9, 9345 (2019). https://doi.org/10.1038/s41598-019-45835-3.
3. Betschart, R.O., Thiéry, A., Aguilera-Garcia, D. et al. Comparison of calling pipelines for whole genome sequencing: an empirical study demonstrating the importance of mapping and alignment. Sci Rep 12, 21502 (2022). https://doi.org/10.1038/s41598-022-26181-3.
4. Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, et al. (2013) Coverage Bias and Sensitivity of Variant Calling for Four Whole-genome Sequencing Technologies. PLOS ONE 8(6): e66621. https://doi.org/10.1371/journal.pone.0066621.
5. Kosugi, S., Momozawa, Y., Liu, X. et al. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol 20, 117 (2019). https://doi.org/10.1186/s13059-019-1720-5.
6. Zook, J., Chapman, B., Wang, J. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol 32, 246–251 (2014). https://doi.org/10.1038/nbt.2835.
7. Zook, J., Catoe, D., McDaniel, J. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 3, 160025 (2016). https://doi.org/10.1038/sdata.2016.25
8. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data; Babraham Institute: Cambridge, UK, 2017.
9. Bushnell, B. BBMap: a fast, accurate, splice-aware aligner. 2014. Available online: https://github.com/BioInfoTools/BBMap.
10. Li, H.; Durbin, H. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009, 25, 1754–1760.
11. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map Format and SAMtools. Bioinformatics 2009, 25, 2078–2079.
12. Broad Institute. Picard Toolkit. 2014. Available online: https://broadinstitute.github.io/picard/.

13. Li, H. A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data. Bioinformatics 2011, 27, 2987–2993.
14. Poplin R. et al. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology 36, 983–987 (2018).
15. Tan, A.; Abecasis, G.R.; Kang, H.M. Unified Representation of Genetic Variants. Bioinformatics 2015, 31, 2202–2204.
16. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29, 308–311 (2001).
17. Vasiliadis Iu.A., Suchalko O.N., Shmitko A.O., Pavlova A.S., Belova V.A., Samitova A.F., Korostin D.O. Program "Exome-GID" for automatic identification of related or paired samples (healthy and tumor tissue) by focused genotype. ROSPATENT. Certificate №2023688418 from 22.12.2023.
18. Sosa MX, Sivakumar IKA, Maragh S, Veeramachaneni V, Hariharan R, et al. (2012) Next-Generation Sequencing of Human Mitochondrial Reference Genomes Uncovers High Heteroplasmy Frequency. PLOS Computational Biology 8(10): e1002737. https://doi.org/10.1371/journal.pcbi.1002737.
19. Weissensteiner H, Forer L, Fuchsberger C, Schöpf B, Kloss-Brandstätter A, Specht G, Kronenberg F, Schönherr S. 2016. mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. Nucleic Acids Res 44: W64–9.