WHOLE GENOME SEQUENCING POPULATION STUDIES

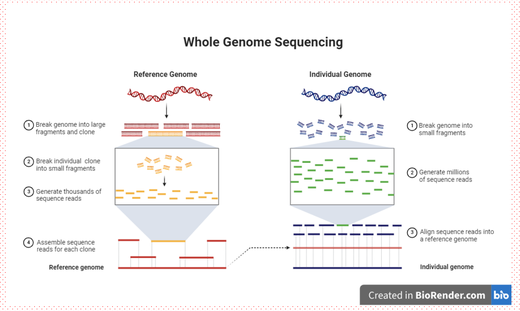
Zuzana Rozankova[1], Patricia Zizkovicova[1], Helena Jurtikova[1,2], Barbora Blumova[1], Veronika Holinkova[2], Jiri Drabek[1,2], Marian Hajduch[1,2,3]

[1] Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacký University Olomouc, Olomouc, the Czech Republic.
[2] Laboratory of Experimental Medicine, University Hospital, Olomouc, the Czech Republic.
[3] Czech Advanced Technology and Research Institute, Institute of Molecular and Translational Medicine, Palacký University Olomouc, Olomouc, the Czech Republic.

Introduction

Massively parallel sequencing, currently the most widely used method for variant detection, is generating large amounts of sequencing data increasingly faster and cheaper than in previous years. Just as the size of data generated by whole genome sequencing (WGS), protein-coding sequences (WES) or regions of interest (panel sequencing) varies, different approaches can result in a diversity of sequencing data generated, which are then aligned to a reference genomic sequence (Fig. 1). The aim of this study was to create a reference genomic database of variants in a healthy Czech population to compare their frequencies with those detected in other populations.

Fig. 1: WGS – Reference and Individual Genome



Material and method

In this genome-wide study, we included a total of 1159 gDNA samples obtained from healthy blood donors of Czech nationality whose parents had to meet the Czech nationality requirement. Criteria for participants to take part in the study were as follows: age between 30 and 55 years, absence of serious genetic disease and both participant´s parents come from the same region of the Czech Republic. Mother and father can come from two different regions if their places of birth are less than 50 km apart as the crow flies. Exclusion criteria to avoid presence of foreign DNA (chimerism) were undergoing an organ, tissue or bone marrow transplant at any time during the participant's lifetime, receipt of a blood transfusion within the last month and ongoing pregnancy. Participant data (including health status and medical history) were voluntarily reported by participants, no reference to medical registries was made. The project was approved by the ethics committee of the university. All participants consented to the processing of their data and samples, which was documented by signing an informed consent form. The input amount of DNA, isolated from peripheral blood leukocytes, for the preparation of the sequencing library using the TruSeq DNA PCR Free kit was 1 µg. All samples were mechanically fragmented to the 350 bp fragment size using Covaris focused ultrasonicator prior to library preparation. TruSeq DNA UD Indexes v0 were used for library preparation. The quality control (QC) of all completed libraries was performed prior to sequencing using a Qubit® 2.0 Fluorometer with Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific), 2100 Bioanalyzer using High Sensitivity DNA kit (Agilent Technologies) and CFX96 Touch Real-Time PCR detection system (Bio-Rad) using KAPA Human Genomic DNA Quantification and QC Kit (Roche Diagnostics). Sequencing was performed on an Illumina NovaSeq 6000 with a read length of 160 bp PE.

Results and discussion

The entire human genome was sequenced using Illumina technology in several sequencing cycles. Minimum coverage was set to 30x at 80% of the GATK GRCh38 reference including contigs. The average total number of reads obtained was 390 million with an average read length of 152 bp and an average coverage per human genome of 39.11, with a minimum coverage of 27.63 and a maximum coverage of 54.37. Bioinformatics analysis was performed using a custom bioinformatics pipeline. In terms of removing poor-quality data, we performed quality control of the raw .fastq files using FastQC and MultiQC. Subsequently, quality-controlled sequencing reads were mapped to the reference genome followed by detection of variants. All variants were then filtered, annotated, and prepared for further analysis and interpretation. From the whole genome data obtained was created the reference genomic database of variants in a healthy Czech population.

Sequence per read metrics and Sequencing quality parameters (Mean Quality Scores, Per Sequence GC content, and Sequence Duplication Levels) have shown satisfactory quality of sequencing data. (Fig. 2, Fig. 3, Fig. 4, Fig. 5).

Fig. 2: Per Read Metrics

Per Read Metrics

READ	CYCLES	YIELD	PROJECTED YIELD	ALIGNED (%)	ERROR RATE (%)	INTENSITY CYCLE 1	%>Q30
Read 1	161	2.04 Tbp	2.04 Tbp	1.13	0.22	1057.26	93.01
Read 2 (I)	8	89.30 Gbp	89.30 Gbp	0.00	0.00	989.66	94.88
Read 3 (I)	8	89.20 Gbp	89.20 Gbp	0.00	0.00	998.71	92.97
Read 4	161	2.04 Tbp	2.04 Tbp	1.13	0.27	690.74	89.16
Non-index Reads Total	322	4.08 Tbp	4.08 Tbp	1.13	0.24	874.00	91.08
Total	**338**	**4.26 Tbp**	**4.26 Tbp**	**1.13**	**0.24**	**934.09**	**91.20**

Fig. 3: FastQC quality output: Mean Quality Scores

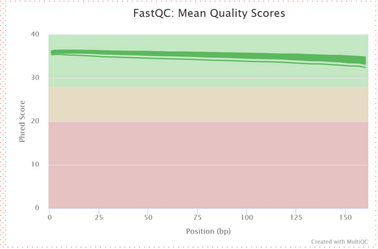


Fig. 4: FastQC quality output: Per Sequence GC Content

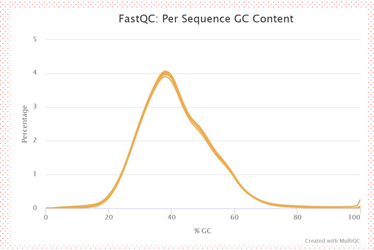
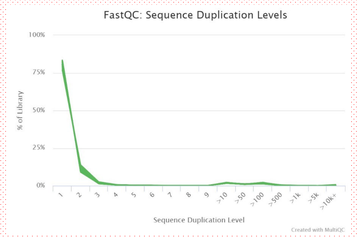


Fig. 5: FastQC quality output: Sequence Duplication Levels



Conclusion

It may be concluded that whole genome sequencing (WGS) technology is a cost-effective solution for analysing chromosomal abnormalities, copy number variations (CNVs), gene fusions, germline variants, loss of heterozygosity (LOH), single nucleotide polymorphisms (SNPs), multiple-nucleotide variants (MNV), small insertions-deletions (indels), somatic variants, structural variants or de novo mutations in the human genome. All genome-wide association studies have confirmed that while common variants are shared by populations worldwide, rarer variants are often restricted to closely related populations. On the one hand, this fact allows us to study the history and demography of our ancestors; on the other hand, it means that many causal variants or prognostic markers may be strictly population-specific, making it impossible to compare data from different unrelated populations. If we do not know the population-specific polymorphisms that are commonly found in the local healthy population, and a variant is found in a disease-related gene, this often leads to difficulties in interpreting such findings. For this reason, it is preferable to create databases collecting data from smaller geographical areas or even from specific countries.

Acknowledgement
This study was supported by the project SALVAGE (OP JAC; reg. no. CZ.02.01.01/00/22_008/0004644) – co-funded by the European Union and by the State Budget of the Czech Republic.