Genet. Res., Camb. (2017), vol. 99, e6. © Cambridge University Press 2017
doi:10.1017/S0016672317000040

Analysis of pathogenic variants from the ClinVar database in healthy people using next-generation sequencing

TAUTVYDAS RANČELIS[1], JUSTAS ARASIMAVIČIUS[1], LAIMA AMBROZAITYTĖ[1], INGRIDA KAVALIAUSKIENĖ[1], INGRIDA DOMARKIENĖ[1], DOVILĖ KARČIAUSKAITĖ[2], ZITA AUŠRELĖ KUČINSKIENĖ[2] AND VAIDUTIS KUČINSKAS[1]*

[1]*Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University, Lithuania*
[2]*Department of Physiology, Biochemistry, Microbiology and Laboratory Medicine, Faculty of Medicine, Vilnius University, Lithuania*

(*Received 30 March 2017; revised 4 July 2017; accepted 25 July 2017*)

Summary

Next-generation sequencing (NGS) became an effective approach for finding novel causative genomic variants of genetic disorders and is increasingly used for diagnostic purposes. Public variant databases that gather data of pathogenic variants are being relied upon as a source for clinical diagnosis. However, research of pathogenic variants using public databases data could be carried out not only in patients, but also in healthy people. This could provide insights into the most common recessive disorders in populations. The study aim was to use NGS and data from the ClinVar database for the identification of pathogenic variants in the exomes of healthy individuals from the Lithuanian population. To achieve this, 96 exomes were sequenced. An average of 42 139 single-nucleotide variants (SNVs) and 2306 short INDELs were found in each individual exome. Pooled data of study exomes provided a total of 243 192 unique SNVs and 31 623 unique short INDELs. Three hundred and twenty-one unique SNVs were classified as pathogenic. Comparison of the European data from the 1000 Genomes Project with our data revealed five pathogenic genomic variants that are inherited in an autosomal recessive pattern and that statistically significantly differ from the European population data.

1. Introduction

Due to technological advances and dramatic sequencing cost reduction, next-generation sequencing (NGS) has been widely used for finding causative genomic variants of genetic disorders. Many laboratories use whole-exome sequencing due to the fact that approximately 85% of all known genetic Mendelian disorders affect protein-coding regions (Gilissen *et al.*, 2012). As exomes comprise approximately only 1% of the whole genome, sequencing the exome instead of the whole genome reduces diagnosis cost and facilitates data analysis.

After acquiring sequencing data, public variant databases, which gather data on known pathogenic genome variants, are being relied upon as a source for clinical diagnosis (Lindor *et al.*, 2017). Too excessive reliance on databases' data, without additional verification, raises concerns with regards to misinterpretation that may harm patients. According to Bell *et al.* (2011), approximately 10% of disease-causing mutations depicted in widely used databases are misinterpreted and such databases should be carefully scrutinized. Nevertheless, public pathogenic variants databases are a beneficial tool for interpretation of sequencing data.

The usage of public databases may also be implemented for pathogenic variants analysis in healthy people. Such research could provide an insight into the most common recessive disorders in a particular population, hence providing possible benefits for diagnostics.

The aim of this study was to use NGS and to apply data from the ClinVar database for identification of pathogenic variants in the exomes of healthy individuals

* Corresponding author: Prof. Vaidutis Kučinskas, Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University, Santariškių 2, LT-08661 Vilnius, Lithuania. E-mail: vaidutis.kucinskas@mf.vu.lt

from the Lithuanian population and to determine frequency differences of pathogenic variants comparing Lithuanian population data and other population data. In order to achieve this, 96 self-reported healthy individuals' exomes were sequenced.

Analysis was also conducted on how well the currently available ClinVar database of pathogenic variants is balanced for healthy individuals' research, since the efficiency of study results is highly dependent on the accuracy of the data in the database (Landrum *et al.*, 2014).

2. Materials and methods

(i) *Samples*

This study is a part of a project called 'Genetic diversity of the population of Lithuania and changes of its genetic structure related with evolution and common diseases' (acronym: LITGEN) (Uktverytė *et al.*, 2013).

Sequencing data of 96 self-reported healthy unrelated individuals (equal male:female ratio) from the Lithuanian population with at least three generations living in Lithuania was used for the analysis. DNA was extracted from venous blood using either the phenol-chloroform method or MagneSil® Genomic, Large Volume System (Promega Corp., USA) on TECAN Freedom EVO® (Tecan, Switzerland).

(ii) *Sequencing*

The 5500 SOLiD™ System (Applied Biosystems; Thermo Fisher Scientific, Inc., USA) was used to sequence the samples. Sequencing was carried out according to the manufacturer's protocols (Thermo Fisher Scientific, Inc., USA) using the SureSelect^{XT} Target Enrichment System (Agilent Technologies, Inc., USA) or the TargetSeq™ Exome Enrichment System (Life Technologies; Thermo Fisher Scientific, Inc., USA). Using the 5500 SOLiD System, 75-bp short-read sequences were generated.

(iii) *Bioinformatic analysis*

The SOLiD System uses a specific ligation-based sequencing strategy and is colour-space encoded. Since previous studies have shown that proprietary software for the SOLiD System is most appropriate for the computational pipeline of data generated by the SOLiD System, LifeScope™ 2·5·1 genomic analysis software was used for mapping to reference genome and variant calling (Pranckevičienė *et al.*, 2015).

Analysis of sequenced data showed that rate of transitions/transversions was 2·2–2·8. These values indicate that obtained data was not generated randomly. In this study 80% of target exons were covered at more than 20X.

To achieve the overview of genomic variance in the sample group representing the Lithuanian population, the Genome Analysis Toolkit's (GATK) Combine Variants software was used to pool all 96 individuals' .vcf files (DePristo *et al.*, 2011). The Integrative Genomics Viewer (IGV) was used for visualization of data (Thorvaldsdóttir *et al.*, 2013). Functional annotation of genomic variants was performed by using ANNOVAR and included frequencies of genomic variants from the 1000 Genomes Project (1000G) and The Exome Aggregation Consortium (ExAC) databases and frequencies of pathogenic variants from the ClinVar database (Wang *et al.*, 2010).

Filtering processes were based on the ClinVar database's overall clinical significance of genome variants. It includes both rare and common variants if they were interpreted as pathogenic.

(iv) *Statistics*

Pathogenic variants with distribution not following Hardy–Weinberg equilibrium, were excluded from this study.

In the comparative frequency analysis of pathogenic variants in the Lithuanian population study group and of other population data (1000G, ExAC), Fisher's exact test was used.

3. Results and discussion

Each individual exome had an average of 42 139 single-nucleotide variants (SNVs) and 2306 short INDELs (up to 19 nucleotide deletions, up to four nucleotide insertions) that differed from the reference genome (hg19). In the 96 Lithuanian exomes that were sequenced, 321 SNVs and 30 short INDELs were identified that were classified as likely pathogenic or pathogenic in the ClinVar database at least by one submitter. Thirteen of them did not following Hardy–Weinberg equilibrium, leaving 308 genome variants in total. An average of 39 SNVs per individual exome were indicated as likely pathogenic or pathogenic by the ClinVar database (Table 1).

In the dataset of this study, 40 genomic variants that were indicated as likely pathogenic or pathogenic by ClinVar had 25% and higher frequencies in the 1000 G and ExAC data. Pathogenic variants with such high frequency in the global population are highly questionable, and therefore these variants were excluded from further analysis. Of all the genomic variants indicated as likely pathogenic or pathogenic, 277 genomic variants had a frequency lower than 25%, and 147 genomic variants had a frequency lower or equal to 1% in the 1000 G and ExAC data (Table 2).

A debatable issue is that there were individuals who had homozygous genotypes for alleles identified as

https://doi.org/10.1017/S0016672317000040 Published online by Cambridge University Press

Table 1. *The total number of genomic variants from 96 exomes and an average number of them in a single exome, together with clinical significance provided by the ClinVar database.*

		Groups of clinical significance by ClinVar			
	Total genomic variants	Non-pathogenic	Likely non-pathogenic	Likely pathogenic	Pathogenic
Genomic variants from 96 Lithuanian exomes	243 192 SNVs 31 623 INDELs	3983 SNVs 85 INDELs	1374 SNVs 14 INDELs	28 SNVs 5 INDELs	280 SNVs 25 INDELs
Genomic variants in average exome	42 139 SNVs 2306 INDELs	748 SNVs 13 INDELs	317 SNVs 3 INDELs	1 SNVs less than 1	39 SNVs less than 1

Table 2. *Statistics of SNVs and short INDELs considered as likely pathogenic or pathogenic in the ClinVar database in self-reported healthy Lithuanian individuals compared to other population data.*

SNV	
Frequency in 1000 G and ExAC <25%	277
Intronic variants	10
Splicing variants	6
Exonic variants	261
Nonsynonymous	250
Synonymous	11
Frequency in 1000 G and ExAC ⩽1%	147
Intronic variants	2
Splicing variants	6
Exonic variants	138
Nonsynonymous	135
Synonymous	3
Short INDEL	
Frequency in 1000 G and ExAC <25%	30
Intronic variants	2
Splicing variants	–
Exonic variants	28
Frameshift variants	25
Nonframeshift variants	3
Frequency in 1000 G and ExAC ⩽1%	24
Intronic variants	0
Splicing variants	–
Exonic variants	24
Frameshift variants	23
Nonframeshift variants	1

pathogenic by ClinVar, meaning that these individuals may have disease symptoms. Since data in this study is acquired from self-reported healthy Lithuanian individuals, a possible explanation is that the pathogenic variant causes a very subtle alteration, or that the phenotype was not determined in detail, or that the variant is incorrectly attributed as pathogenic. For some pathogenic variants, this homozygous state is seen both in our data and in the data of large-scale population studies (Lek *et al.*, 2016).

Pathogenic variants identified in Lithuanians were grouped according to medical disease classification 'International Statistical Classification of Diseases and Related Health Problems' 10th revision (ICD-10). In the Lithuanian individuals, the most common diseases associated with the pathogenic variants studied are endocrine, nutritional and metabolic diseases (which account for 30·5% of diseases), diseases of the blood and blood-forming organs, including disorders of the immune mechanism (which account for 15·2% of diseases), and congenital malformations, congenital deformations and chromosomal abnormalities (which account for 14·3% of diseases) (Table 3).

A relatively high number (7·6%) of diseases were of ophthalmic origin. Further analysis showed that several pathogenic genome variants related to ophthalmic diseases have a higher frequency in Lithuanian individuals.

For further comparative analysis, if SNVs or short INDELs in the Lithuanian individuals were assigned as pathogenic, then frequency comparison with other population data from the ExAC and 1000 G projects was performed. Comparison with ExAC data without psychiatric cohorts (from 45 376 unrelated individuals) showed that as many as 95 likely pathogenic or pathogenic variants in our study differed in a statistically significant manner from ExAC data. To acquire more Lithuanian-specific pathogenic variants, comparison with the 1000 G European data was performed for all assigned pathogenic variants (1000 Genomes Project Consortium *et al.*, 2012).

Pathogenic variants can have a very low frequency and even in large-scale population studies can appear in small numbers. Since a small number of alleles cannot impartially represent a statistically significant difference in the present study group or other populations, the criteria of a minimum of four alleles was set for further analysis. Out of an overall 30 unique short INDELs, 25 did not pass this requirement and the frequency of the other five INDELs demonstrated no statistically significant difference from the 1000 G European data. The SNV comparison of pathogenic variant frequencies in studied individuals with frequencies of pathogenic variants from the genomic data of Europeans identified five pathogenic genomic variants that are inherited in an autosomal recessive

https://doi.org/10.1017/S0016672317000040 Published online by Cambridge University Press

Table 3. *Classification of diseases potentially related to pathogenic variants in self-reported healthy Lithuanian individuals.*

Disease classification	ICD-10 code group	Frequency in the Lithuanian group
Endocrine, nutritional and metabolic diseases	E	30·5%
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	D	15·2%
Congenital malformations, congenital deformations and chromosomal abnormalities	Q	14·3%
Diseases of the eye and adnexa	H	7·6%
Diseases of the nervous system	G	6·7%
Diseases of the circulatory system	I	6·7%
Diseases of the digestive system	K	4·8%
Malignant neoplasms	C	2·9%
Symptoms, signs, and abnormal clinical and laboratory findings not classified elsewhere	R	2·9%
Infectious diseases	B	1·8%
Diseases of the musculoskeletal system and connective tissue	M	1·8%
Diseases of the genitourinary system	N	1·8%
Diseases of the ear and mastoid process	H	1·0%
Diseases of the respiratory system	J	1·0%
Complications of surgical and medical care not classified elsewhere	T	1·0%

Table 4. *dbSNP database codes of pathogenic variants which have autosomal recessive inheritance and which frequencies statistically significantly differ in comparison of Lithuanian individuals' data with 1000 Genomes Project's European data.*

dbSNP	Gene	Disorder	1000G	LTU	Fisher
Genomic variants with statistically significantly higher frequencies in Lithuanian individuals					
rs1800553	*ABCA4*	Stargardt disease	0·40	2·55	$8{\cdot}01 \times 10^{-3}$
rs142181517	*PHYKPL*	Phosphohydroxy lysinuria	0·40	3·57	$5{\cdot}21 \times 10^{-4}$
rs113298164	*LIPC*	Hepatic lipase deficiency	0·50	2·04	0·0447
rs34526199	*AMPD1*	Muscle AMP deaminase deficiency	3·68	7·14	0·053
rs104895094	*MEFV*	Familial Mediterranean Fever	0·55	6·63	$1{\cdot}16 \times 10^{-6}$

1000G: genomic variants' frequency in 1000 Genomes Project, dbSNP: data from build 149 dbSNP database; Fisher: Fisher's exact test; **LTU:** genomic variants' frequency in LITGEN project.

manner and have no conflicting interpretation in the ClinVar database (Table 4).

In addition, another seven genome variants, which are inherited in an autosomal recessive manner, show a statistically significant difference; however, they have conflicting interpretation in the ClinVar database.

4. Conclusions

The study results represent statistically significant differences in frequencies of genomic variants between individuals from the Lithuanian population and other populations.

When the data of the present study was compared with all ExAC project data, a statistically significant difference was observed for 95 out of 308 likely pathogenic or pathogenic variants, but most of them correlated with the frequencies of European data. Comparison with the European data of the 1000 Genomes Project revealed five statistically significant pathogenic variants that differed from the European population data and have no conflicting interpretation in the ClinVar database.

The study showed that whole-exome sequencing and analysis of the general population is an effective way to find pathogenic variants with statistically significant differences in a particular population even if the cohort studied is relatively small. This could be valuable information for genetic counselling and may benefit clinical diagnosis by focusing on the specific variants that are more frequent in a particular population.

Based on the present study data, ClinVar is currently the best freely available database of genomic variants of different clinical significance. A considerable

https://doi.org/10.1017/S0016672317000040 Published online by Cambridge University Press

amount of the variants classified as pathogenic in ClinVar have a high frequency in 1000 G and ExAC. We observed a similar pattern in LITGEN data.

Another matter of concern is that there were individuals who had homozygous genotypes for alleles identified as pathogenic, thus cautious interpretation of the ClinVar data for pathogenic variants should be undertaken by researchers and medical specialists.

The research leading to these results is part of the LITGEN project (VP1-3·1-ŠMM-07-K-01-013) and was funded by the European Social Fund under the Global Grant measure.

Declaration of interest

The authors declare no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Ethical approval

All procedures performed in this study involving human participants were in accordance with the ethical standards of the Vilnius Regional Research Ethics Committee (No. 158200-05-329-79. date: 2011-05-03) and with the 1964 Helsinki declaration and its later amendments.

References

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., Depristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T. & McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65.

Bell, C. J., Dinwiddie, D. L., Miller, N. A., Hateley, S. L., Ganusova, E. E., Mudge, J., Langley, R. J., Zhang, L., Lee, C. C., Schilkey, F. D., Sheth, V., Woodward, J. E., Peckham, H. E., Schroth, G. P., Kim, R. W. & Kingsmore, S. F. (2011). Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Science Translational Medicine* **3**, 65ra4.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D. & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491–498.

Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. (2012). Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics* **20**, 490–497.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M. & Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research* **42**, D980–D985.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Posthuma, D. & Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291.

Lindor, N. M., Thibodeau, S. N. & Burke, W. (2017). Whole-genome sequencing in healthy people. *Mayo Clinic Proceedings* **92**, 159–172.

Pranckevičienė, E., Rancelis, T., Pranculis, A. & Kucinskas, V. (2015). Challenges in exome analysis by LifeScope and its alternative computational pipelines. *BMC Research Notes* **8**, 421.

Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**, 178–192.

Uktverytė, I., Meškienė, R., Ambrozaitytė, L., Domarkienė, I., Pranculis, A., Burokienė, N., Coj, A., Mažeikienė, A., Kasiulevičius, V., Kučinskienė, Z. A. & Kučinskas, V. (2013). LITGEN – revealing genetic structure of the population of Lithuania. *European Journal of Human Genetics* 21 Suppl. **2**, 394.

Wang, K., Li, M. & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164.