CLARIFICATION OF THE STATUS OF SOME MUTATIONS CONSIDERED PATHOGENIC, BY HARMLESS MUTATIONS ATTRIBUTES

Borisevich DI[1,2], Shatalova LV[1], Korostin DO[1,3] ✉, Ilinsky VV[1,3]

[1] Bioinformatics Data Processing Department,
Genotek Ltd., Moscow, Russia

[2] Lomonosov Moscow State University, Moscow, Russia

[3] The Core Facilities Center "Genetic Polymorphism"
Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia

Prediction of mutation pathogenicity and its effect on the phenotype is an important task of modern bioinformatics. This task is particularly difficult in regard to single nucleotide polymorphisms, as their effect is very hard to predict. Information on pathogenic mutations is provided by curated databases such as Online Mendelian Inheritance in Man (OMIM) and The Human Gene Mutation Database (HGMD) which include data from experimental works. However, as different authors interpret the term "mutation pathogenicity" differently, it is necessary to double-check data before using them. We have assessed HGMD database quality using the most common bioinformatic tools, namely, snpEff, polyphen2 and SIFT. Our study relied on the characteristics specific for harmless mutations: high frequency in a population, weak effect on amino acid sequence of a protein, low pathogenicity as computed by the utilities used in the study. As a result, we have identified clearly harmless variants among those in the mutation database, as well as ambiguous ones in which a mutation type depends on characteristics and tools used for the analysis.

Keywords: human genetics, high-throughput sequencing, pathogenicity, population analysis, search for mutations

✉ **Correspondence should be addressed:** Dmitriy Korostin
ul. Gubkina, d. 3, Moscow, Russia, 119991; d.korostin@gmail.com

Recieved: 10.02.2016 **Accepted:** 19.02.2016

УТОЧНЕНИЕ СТАТУСА НЕКОТОРЫХ МУТАЦИЙ, СЧИТАЮЩИХСЯ ПАТОГЕННЫМИ, С ПОМОЩЬЮ ПРИЗНАКОВ БЕЗВРЕДНЫХ МУТАЦИЙ

Д. И. Борисевич[1,2], Л. В. Шаталова[1], Д. О. Коростин[1,3] ✉, В. В. Ильинский[1,3]

[1] Отдел биоинформатической обработки данных,
ООО «Генотек», Москва

[2] Московский государственный университет имени М. В. Ломоносова, Москва

[3] Центр коллективного пользования отдела биологических наук РАН «Генетический полиморфизм»,
Институт общей генетики имени Н. И. Вавилова РАН, Москва

Важной задачей современной биоинформатики является предсказание патогенности мутации и ее влияния на фенотип. Она особенно трудна для однонуклеотидных полиморфизмов, чей эффект сложнее всего предсказать. Патогенные мутации берут из курируемых баз данных, таких как Online Mendelian Inheritance in Man (OMIM) и The Human Gene Mutation Database (HGMD), куда включают данные из экспериментальных статей. Однако поскольку различные авторы вкладывают разный смысл в понятие «патогенность мутации», необходимо контролировать данные баз перед их использованием. Мы проанализировали качество данных базы HGMD с помощью наиболее часто используемых биоинформатических инструментов: snpEff, polyphen2 и SIFT. В исследовании мы опирались на признаки, характерные для безвредных мутаций: высокую частоту в популяции, слабое влияние на аминокислотную последовательность белка, низкую патогенность по оценке вычислительных методов. В результате среди мутаций базы нами выявлены однозначно безвредные варианты, а также варианты со спорным значением, для которых тип мутации зависит от используемых для анализа признаков и инструментов.

Ключевые слова: генетика человека, высокопроизводительное секвенирование, патогенность, популяционный анализ, поиск мутаций

✉ **Для корреспонденции:** Дмитрий Олегович Коростин
119991, г. Москва, ул. Губкина, д. 3; d.korostin@gmail.com

Статья получена: 10.02.2016 **Статья принята в печать:** 19.02.2016

The impact of single nucleotide polymorphisms (SNP) on the phenotype is hard to predict. Currently existing tools for predicting mutation pathogenicity have a number of flaws, such as low sensitivity and specificity of no more than 75–80 % for SNP. Besides, they often do not annotate insertions and deletions [1–3].

Pathogenic mutations described in experimental articles are collected into databases, such as the Online Mendelian Inheritance in Man database (OMIM, [4]) and the Human Gene Mutation Database (HGMD [5]). However, the term *pathogenicity* can be interpreted widely; there is no unanimous opinion on what it implies. As a result, different approaches are applied while selecting mutations for their inclusion in a database; thus, the data in different databases are not the same and need rectification.

To identify non-pathogenic mutations, their indirect indicators are often used, such as allele frequency in a population and the effect on the amino acid sequence of a protein. With new data coming into sight, these indicators can help us understand how the existing databases can be improved. Knowing that mutations described as pathogenic meet the criteria for non-pathogenic variants is important for the practical usage of the data derived from these databases. This knowledge can help us understand why certain genetic variants affect the phenotype while others do not.

For scientists who rely on HGMD in their research it may not be obvious that apart from clearly deleterious mutations, it currently includes harmless ones assessed as pathogenic. Within the framework of this study, the pathogenicity of mutations included in HGMD was evaluated using bioinformatic tools. Allele frequencies annotated in HGMD were compared to those from Exome Aggregation Consortium 0.3 [6]; the effect of HGMD mutations on the amino acid sequence of proteins was analyzed, and their pathogenicity was predicted using the most common bioinformatic tools: snpEff, PolyPhen-2 and SIFT.

METHODS

A public version of HGMD (of the fourth quarter of 2014) was used as a source of pathogenic mutations. It contained 73,208 mutations. Their allele frequencies were calculated using snpEff 4.0. The obtained data were compared to the allele frequencies from Exome Aggregation Consortium 0.3 that included whole exome and whole genome sequencing data from 60,706 samples of unrelated patients. ExAC provides allele frequency data on six populations: African, Latino, East Asian, South Asian, Finnish and European (non-Finnish). All unidentified samples are grouped as “Other”. When we used the database, the number of genotyped samples for each annotated mutation varied in different populations, from about 500 for “Other” to 30,000 for Europeans. Allele frequencies were compared using bcftools [7].

HGMD mutations affecting the amino acid sequence of proteins were identified using snpEff 4.0 [8]. A possible level of pathogenicity was predicted using PolyPhen-2 and SIFT utilities. These utilities are standard tools for predicting mutation pathogenicity; neither of them used HGMD data as a training set.

RESULTS

snpEff annotation

Mutations obtained from HGMD were annotated by snpEff, frequencies of each mutation type were established according to snpEff classification. We have found that in many cases mutations have more than one prediction, meaning they can refer to various types at the same time. It usually happens when a mutation is located within the gene and the adjacent genes are used for its annotation. We have filtered variants belonging to more than one type and selected those with the most conspicuous impact according to the algorithm suggested by snpEff developers (see the table below) [8].

Annotation with ExAC

18,159 (25 %) mutations present in HGMD are described in ExAC.

Results obtained by PolyPhen-2 and SIFT

We have predicted mutation pathogenicity using PolyPhen-2 and SIFT utilities. PolyPhen-2 uses two models for pathogenicity prediction: HumDiv and HumVar. According to the developers’ description, HumVar predicts Mendelian diseases better, while HumDiv is more efficient with complex phenotypes and mildly deleterious alleles [9]. We have chosen HumDiv model to use a wider pathogenicity definition. Threshold for cutting off pathogenic and possibly pathogenic variants was set by default.

PolyPhen-2 annotated 52,248 mutations, 39,032 (72 %) of them were identified as pathogenic and 6,220 (11 %) as possibly pathogenic. SIFT utility analyzed 53,097 mutations with 34,638 (65 %) identified as pathogenic and 4,358 (8 %) as possibly pathogenic (with low probability). Both utilities recognized the variants submitted to the database as pathogenic in 70–80 % cases, which corresponds to their expected performance [2, 3].

DISCUSSION

Using ExAC database as a resource containing data on allele frequency

Technical description of ExAC has not been released yet, but the database is known to include data from both population genetic studies and sequencing projects describing the samples of patients with various diseases. We believe that such projects use less samples compared to population genetic research works, and their effect on the resulting frequency must be negligible, especially if samples of a large number of individuals have been analyzed in population genetic studies. That is why our analysis did not cover mutations that had been genotyped in a few individuals only. That being said, we believe that ExAC can certainly be used to estimate the frequencies in such studies as ours. The developers of this database claim that it can be used as a reference set of allele frequencies for disease studies.

Presence of synonymous mutations in HGMD

95 % of all mutations obtained from HGMD were distributed by snpEff in two groups: missense mutations and nonsense mutations. However, about 2.5 % of mutations were identified as synonymous (see the table). Although the pathogenicity of synonymous variants has been described in literature, in most cases synonymous mutations are considered harmless. We focused on this group as a group of variants with the most disputable pathogenicity. PolyPhen-2 utility does not perform the pathogenicity assessment of synonymous mutations because it relies on the effect of a mutation on the protein amino acid sequence. SIFT utility allows for the assessment of the synonymous mutation pathogenicity; it identified only 4 out of 1,793 synonymous mutations as pathogenic. It is highly probable that the rest of 1,789 mutations (~2.5 % of all mutations in HGMD) are not pathogenic because they do not have any other signs of pathogenicity.

Number of the most important mutations obtained from HGMD and predicted by snpEff

Type*	Number of mutations	Type *	Number of mutations
missense_variant	56136	sequence_feature	66
stop_gained	13513	initiator_codon_variant	61
synonymous_variant	1793	intron_variant	54
start_lost	465	non_coding_exon_variant	44
3_prime_UTR_variant	363	splice_donor_variant	39
downstream_gene_variant	245	splice_acceptor_variant	23
upstream_gene_variant	162	stop_retained_variant	4
stop_lost	136	5_prime_UTR_variant	3
splice_region_variant	99	intergenic_region	2

*Names are given as they appear in snpEff. missense_variant – missense mutations; stop_gained – nonsense mutations; synonymous_variant - synonymous mutations; start_lost – a codon variant that changes at least one base of the canonical start codon; 3_prime_UTR_variant – a UTR variant of the 3' UTR; downstream_gene_variant – a sequence variant located 3' of a gene. upstream_gene_variant - a sequence variant located 5' of a gene; stop_lost – a sequence variant where at least one base of the terminator codon (stop) is changed resulting in an elongated transcript; splice_region_variant – a sequence variant in which a change has occurred within the region; of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron; sequence_feature – a sequence variant within any region initiator_codon_variant – a codon variant that changes at least one base of the first codon of a transcript; intron_variant – a transcript variant occurring within an intron non_coding_exon_variant – a sequence variant that changes non-coding exon sequence of a noncoding transcript; splice_donor_variant – a splice variant that changes the 2 base pair region at the 5' end of an intron; splice_acceptor_variant - a splice variant that changes the 2 base region at the 3' end of an intron; stop_retained_variant – a sequence variant where at least one base in the terminator codon is changed, but the terminator remains; 5_prime_UTR_variant - a UTR variant of the 5' UTR; intergenic_region – a region containing or overlapping no genes that is bounded on either side by a gene, or bounded by a gene and the end of the chromosome.

Analysis of synonymous pathogenic mutations in HGMD

Only one of the four synonymous mutations in HGMD identified as pathogenic by SIFT utility is described in dbSNP [10]. It is NM_005228.3:c.2361G>A (NP_005219.2:p.Gln787=) mutation with rsid *rs1050171*. According to Zhang et al. [11], this mutation is associated with lung cancer; its molecular mechanism of action has not been identified yet. The frequency of the alternative ("mutant") allele A is about 43 %, according to the "1000 genomes" project data presented in dbSNP. The ClinVar database [12] defines this SNP as benign [12]. The reasons for SIFT classifying this mutation as pathogenic are probably related to the conservative position where the mutation occurred. It is located at codon position 3 that is usually less conservative than positions 1 and 2, and gets a lower score. However, for this mutation the PhyloP Vertebrate evolutionary conservation score obtained from UCSC Genome Browser [14], combined with the scores of positions 1 and 2 of adjacent codons, is much higher than the score of other third codon position nucleotides, which is indicative of high conservation of the nucleotide of interest.

After all, the true nature of this mutation is hard to identify. On the one hand, there is evidence that this mutation is non-pathogenic, such as the data from ClinVar database, its synonymous type, the high frequency of the allele variants in the population. On the other hand, the results of prediction using SIFT utility in HGMD and the high evolutionary conservation suggest the pathogenicity of this variant. This example illustrates the difficulty of mutation pathogenicity prediction: even manual analysis cannot provide the unambiguous interpretation of the results, because the mutation type depends on the choice of a tool for analysis.

Variants with a mutation present in a heterozygote only

To analyze the mutations absent in the samples in the homozygous state, we have chosen four mutations, each being present in a heterozygote in more than 75 % of samples and in a homozygote in less than 5 % of samples (according to the ExAC data):

1. chr1:1650845G>A (*rs1059831*, gene *CDK11A*, HGMD phenotype: associated with type 2 diabetes) [14],
2. chr2:112614429G>A (*rs72936240*, gene *ANAPC1*, HGMD phenotype: protein deficit associated with the risk of cancer) [15],
3. chr7:142458451A>T (*rs111033566*, gene *PRSS1*, HGMD phenotype: hereditary pancreatitis) [16],
4. chr17:7197581G>T (*rs189257850*, gene *YBX2*, HGMD phenotype: associated with male infertility) [17].

Homozygous variants 2 and 3 have never been present in any population, homozygous variant 1 has been found in only one out of 8,209 samples in the South Asian population. Strangely, for variant 4 only 203 samples have been genotyped, while for variants 1–3 about 60,000 samples have been genotyped. For variant 4 only one individual out of 52 in the East Asian population has been described as homozygous and 13 individuals out of 62 have been described as homozygous in the Latin American population.

These mutations are mainly found in heterozygotes, which can be explained by the fact that they cause death or at least cannot be inherited. Based on the phenotype analysis, variants 2 and 4 can be excluded as heterozygous because of early death or infertility of their carriers. Variant 4 is the most interesting one, but it is the only variant that has not been genotyped widely. It is difficult to understand why this mutation is highly frequent in one of the populations and why the number of individuals analyzed in this population is so low. Because the number of the individuals analyzed is low, those data have been possibly obtained by analyzing diseased individuals (see the description of ExAC specifics above), so no predictions for this variant are possible. Variant 2 can be described as lethal in the homozygous state. We make a supposition that although it is not obvious that variants 1 and 3 are lethal, the existent data prove that these mutations cause death or infertility in homozygotes.

CONCLUSIONS

Assessing mutation pathogenicity is a difficult task. Sometimes neither automatic nor manual analysis can classify it as clearly pathogenic or harmless. However, in the absence of

experimental data on transgenic organisms with a mutation of interest, the existing databases can still be used for pathogenicity analysis, but one should use them carefully. Automatic use of those databases is restricted by the quality of data presented there. It is important to manually check if the mutations described in experimental works are pathogenic, especially if the claims of their pathogenicity do not correspond to the database prediction.

References

1. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012; 7 (10): e46688.
2. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 2009; 4 (7): 1073–81.
3. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat. Methods. 2010; 7 (4): 248–9.
4. Online Mendelian Inheritance in Man, OMIM [Internet]. Baltimore (MD): McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University. c1996–2016. [cited 2016 Feb]. Available from: http://omim.org/.
5. The Human Gene Mutation Database, HGMD [Internet]. Cardiff (UK): Cardiff University. c2015 – [cited 2016 Feb]. Available from: http://www.hgmd.cf.ac.uk/ac/index.php.
6. Exome Aggregation Consortium (ExAC) [Internet]. Cambridge (MA). [updated 2016 Jan 17, cited 2016 Feb]. Available from: http://exac.broadinstitute.org/.
7. Bcftools. Available from: http://samtools.github.io/bcftools/.
8. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012; 6 (2): 80–92.
9. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. 2013; Chapter 7: Unit 7.20.
10. dbSNP Short Genetic Variants [Internet]. Available from: http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=1050171.
11. Zhang W, Stabile LP, Keohavong P, Romkes M, Grandis JR, Traynor AM, et al. Mutation and polymorphism in the EGFR-TK domain associated with lung cancer. J. Thorac. Oncol. 2006; 1 (7): 635–47.
12. ClinVar [Internet]. Available from: http://www.ncbi.nlm.nih.gov/clinvar/variation/45271/.
NM_005228.3(EGFR):c.2361G>A (p.Gln787=) Simple - Variation
13. Report - ClinVar – NCBI. Available from: http://www.ncbi.nlm.nih.gov/clinvar/variation/45271/.
14. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010; 20 (1): 110–21.
15. Li Y, Wu G, Zuo J, Gao J, Chang Y, Fang F. Genetic variations of the CDC2L2 gene are associated with type 2 diabetes in a Han Chinese cohort. Diabetes. Metab. Res. Rev. 2007; 23 (6): 455–61.
16. He M-L, Chen Y, Chen Q, He Y, Zhao J, Wang J, et al. Multiple gene dysfunctions lead to high cancer-susceptibility: evidences from a whole-exome sequencing study. Am. J. Cancer Res. 2011; 1 (4): 562–73.
17. Pfützer R, Myers E, Applebaum-Shapiro S, Finch R, Ellis I, Neoptolemos J, et al. Novel cationic trypsinogen (PRSS1) N29T and R122C mutations cause autosomal dominant hereditary pancreatitis. Gut. 2002; 50 (2): 271–2.
18. Hammoud S, Emery BR, Dunn D, Weiss RB, Carrell DT. Sequence alterations in the YBX2 gene are associated with male factor infertility. Fertil. Steril. 2009; 91 (4): 1090–5.

Литература

1. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012; 7 (10): e46688.
2. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 2009; 4 (7): 1073–81.
3. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat. Methods. 2010; 7 (4): 248–9.
4. Online Mendelian Inheritance in Man, OMIM [Интернет]. Baltimore (MD): McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University. c1996–2016. [процитировано в феврале 2016 г.]. Доступно по ссылке: http://omim.org/.
5. The Human Gene Mutation Database, HGMD [Интернет]. Cardiff (UK): Cardiff University. c2015 – [дата обращения: февраль 2016 г.]. Доступно по ссылке: http://www.hgmd.cf.ac.uk/ac/index.php.
6. Exome Aggregation Consortium (ExAC) [Интернет]. Cambridge (MA). [обновлено 17 января 2016 г., процитировано в феврале 2016 г.]. Доступно по ссылке: http://exac.broadinstitute.org/.
7. Bcftools. Доступно по ссылке: http://samtools.github.io/bcftools/.
8. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012; 6 (2): 80–92.
9. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. 2013; Chapter 7: Unit 7.20.
10. dbSNP Short Genetic Variants [Интернет]. Доступно поссылке: http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=1050171.
11. Zhang W, Stabile LP, Keohavong P, Romkes M, Grandis JR, Traynor AM, et al. Mutation and polymorphism in the EGFR-TK domain associated with lung cancer. J. Thorac. Oncol. 2006; 1 (7): 635–47.
12. ClinVar [Интернет]. Доступно по ссылке: http://www.ncbi.nlm.nih.gov/clinvar/variation/45271/.
13. NM_005228.3(EGFR):c.2361G>A (p.Gln787=) Simple - Variation Report - ClinVar – NCBI. Доступно по ссылке: http://www.ncbi.nlm.nih.gov/clinvar/variation/45271/.
14. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010; 20 (1): 110–21.
15. Li Y, Wu G, Zuo J, Gao J, Chang Y, Fang F. Genetic variations of the CDC2L2 gene are associated with type 2 diabetes in a Han Chinese cohort. Diabetes. Metab. Res. Rev. 2007; 23 (6): 455–61.
16. He M-L, Chen Y, Chen Q, He Y, Zhao J, Wang J, et al. Multiple gene dysfunctions lead to high cancer-susceptibility: evidences from a whole-exome sequencing study. Am. J. Cancer Res. 2011; 1 (4): 562–73.
17. Pfützer R, Myers E, Applebaum-Shapiro S, Finch R, Ellis I, Neoptolemos J, et al. Novel cationic trypsinogen (PRSS1) N29T and R122C mutations cause autosomal dominant hereditary pancreatitis. Gut. 2002; 50 (2): 271–2.
18. Hammoud S, Emery BR, Dunn D, Weiss RB, Carrell DT. Sequence alterations in the YBX2 gene are associated with male factor infertility. Fertil. Steril. 2009; 91 (4): 1090–5.