이학석사 학위논문

RarePedia: A tool that integrates scattered information about rare-damaging variants

레어피디아 : 희귀 손상 변이에 대한 산재된 정보를 통합하는 도구

2018년 8월

서울대학교 대학원
생물정보학 협동과정
송 유 림

ABSTRACT

Finding causal factors of various diseases, whether they be environmental factors, stress, aging, and etc. has been the focus of many researchers around the globe. As advancements in science and technology were made, many disease-related genes and mechanisms have been discovered. With the development of DNA sequencing techniques, the sequences of disease-associated genes and specific disease-related genetic variants are being revealed. However, variants that occur at very low frequencies are often ignored, seemingly because there is little known information about these rare variants, which in turn makes rare variant analyses difficult without increasing the sample size.

RarePedia was designed to make a unified collection of information about rare variants. Furthermore, it focuses on deleterious variants that are expected to be related to diseases. RarePedia is another way to use Next Generation Sequencing(NGS) data, and a helpful tool to see organized information previously scattered across many sources.

The ultimate goal of RarePedia is the accumulation of information through additional updates, whenever there is new information about rare and deleterious variants. It can be a way to organize information about rare and deleterious variants that have not been organized systematically.

Keyword : rare-damaging variant, gene, next generation sequencing(NGS), methodology

Student number : 2016-20463

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Finding causal factors of various diseases, whether they be environmental factors, stress, aging, and etc. has been the focus of many researchers around the globe. Several studies have shown that when identical twins are exposed to different environments, disease my or may not appear despite being genetically identical. As the field of genetics developed, it has been found that the reason for this is due to structural changes in DNA that occurs after exposure to a specific environment[1]. Likewise, research efforts on the relationship between disease and stress or aging are steadily increasing.

Furthermore, the underlying mechanisms of many diseases have been identified, and the genes responsible for the diseases are now being discovered. Methods for DNA sequencing have been developed, which allows researchers to identify sequence composition of genes. As a result, it has been confirmed that some mutations occur at significantly high frequencies in patient groups. On the contrary, these mutations are found at very low frequencies in the general population, so it is possible to say that these mutations cause the respective diseases.

These mutations were found to be the result of studies using populations with specific diseases. They were known to be rare, but only because they have been confirmed to appear in patients at a high frequency. However, mutations found in a small number of patients, in other words, rare variants found in individuals who do not have the characteristics of the group, tend to be ignored. This is because rare variants are not only difficult to analyze without increasing the sample size, but also because it is hard to identify what function and role they actually have[2][3]. In other words, there are many difficulties in asserting that there is an association between a rare variant and a particular phenotype, which is why these variants can not be included in many analyses.

RarePedia is designed to integrate and accumulate various information about rare variants that are found at a very low frequency. RarePedia is a tool that generates pages that contain information about rare-damaging variants of an individual through annotation and filtering steps, using input of individual genetic information obtained through next generation sequencing. It parses not only the information that exist in annotation databases, but also the information that is provided in other sources, and presents information on genes and the diseases associated with these variants. It provides a collection of information about rare-damaging variants that are expected to be disregarded because they are found at very low frequencies and are expected to be associated with protein damage or disease.

2. DATA

2.1. Next Generation Sequencing(NGS) Data

As our input Next Generation Sequencing(NGS) data, we used public data provided by 1000 Genomes Project and Alzeheimer's Disease Sequencing Project(ADSP).

2.1.1. 1000 Genomes Project data

The 1000 Genomes Project is an international project which aimed to analyze the genomes of 1000 people composed of various races. Currently, a database of 2,504 genome data from 27 races, is available for anyone to use. Samples analyzed in this project have no phenotype information. However, when providing samples, all providers declared themselves healthy, so data from the 1000 Genomes Project are considered to be healthy individuals. In many studies, this public database has been used for validation and analysis[4].

In RarePedia, we first looked at 2,504 sample data to see how many rare-damaging variants exist in individuals.

2.1.2. Alzeheimer's Disease Sequencing Project(ADSP) data

In addition to the sequencing data provided by the 1000 Genomes Project, the distribution of rare-damaging variants was also confirmed using data provided by the Alzheimer's Disease Sequencing

Project(ADSP).

Alzheimer's disease(AD) can be classified into early-onset Alzheimer's disease(EOAD) and late-onset Alzheimer's disease(LOAD) according to the timing of the onset. ADSP provides a total of 10,913 whole-exome sequencing data for 5,777 AD patients and 5,136 non-AD samples[5]. The statistics of 651 EOAD patients that onset before 65 years of age were obtained. The definition of EOAD patients is defined as AD patients with onset before 65 years of age.

If the sequencing data provided by the 1000 Genomes Project targeted healthy public, the data provided by ADSP focused on AD patients. We used the ADSP sequencing data, assuming that there may be rare-damaging variants that were not found in the 1000 Genomes Project data.

2.2. Databases

2.2.1. Databases prepared by ANNOVAR developers

ANNOVAR(ANNOtate VARiation) is a quick and easy tool that helps annotation of high-throughput sequencing data[6]. Annotated information thorough ANNOVAR is used for variant filtering. The databases used are dbnsfp30a, 1000g2015aug_all, knownGene and clinvar_20160302 (See TABLE 1).

dbnsfp30a is a dataset for functional prediction of variants. It includes many kinds of predictive scores such as SIFT, PolyPhen-2, CADD, GERP++, and so on.

1000g2015aug_all is latest 1000 Genomes Project dataset with allele frequencies in six populations including ALL, AFR(African), AMR(Admixed American), EAS(East Asian), EUR(European), SAS(South Asian).

knownGene is a table from UCSC Genome Browser. In RarePedia, it was used to annotate the name of the gene containing the variant.

2.2.2. Resources used by Oncotator

Oncotator is a tool developed for cancer research that annotates information from a variety of resources[7] (See TABLE 2). Oncotator annotates information for the current version of RarePedia. However, since a large amount of information is annotated, only necessary information is selectively used.

2.2.3. Additional publicly available resources

RarePedia also includes additional information from some publicly available resources (See TABLE 3).

Additional resources used in the “Gene Section” are PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) and GeneRIF (https://www.ncbi.nlm.nih.gov/gene/about-generif). PubMed is a free resource that comprises citations and abstracts for biomedical literature indexed in MEDLINE database[8]. With PubMed, information from life science journals and online books, including abstracts as well as full-text content re available[9]. RarePedia used PubMed to obtain a lst of articles related to genes containing a rare-damaging variant.

GeneRIF(Gene Reference Into Function) provides a simple statement about the function of a gene. It was used for annotation of brief information related to the gene[10].

In the “Transcript Section”, Expression Atlas (https://www.ebi.ac.uk/gxa/home) and The Human Protein Atlas (https://www.proteinatlas.org/) were used to compare the degree of gene expression. Expression Atlas provides information about gene and protein expression in different species and contexts[11]. RarePedia was used to compare the degree of gene expression in each organ revealed by several experiments.

The Human Protein Atlas has provided additional information to get the level of protein expression. It is used to show the distribution of protein in all major tissues and organs of the human body or to show the degree of protein expression in cancer patients[12][13].

The “Protein Section” uses three external resources. UniProt (http://www.uniprot.org/) is a consortium that unifies information on proteins from literatures or different databases[14]. In addition to the information annotated by the Oncotator, it was used to obtain additional information about the protein.

InterPro (http://www.ebi.ac.uk/interpro/) is a database that is used to classify protein families and to locate important domains or sites, using amino acid sequences in proteins[15]. It was used to obtain information about the family, subfamily, and domain of proteins translated from a gene.

Ensembl (http://grch37.ensembl.org/index.html) provides many parts for the analysis of not only humans but also many species of genomes. It effectively visualizes a very large amount of information from various sources and was used to provide domain information of a protein[16].

One gene is involved in several pathways in the body, and these pathway information are provided through different databases such as KEGG, Reactome pathway, and Sino Biological. In the “Pathway Section”, GeneCards(www.genecards.org/) was used to provide known information on genes involved in pathways[17][18].

Lastly, In the“Phenotype Section”, DisGeNET(www.disgenet.org/) database was used to provide information on diseases known to be associated with genes[19]. The current version of DisGeNET(v5.0) contains 561,119 gene-disease associations(GDAs), between 17,074 genes and 20,370 diseases, disorders, traits, and clinical or abnormal human phenotypes, and 135,588 variant-disease associations(VDAs), between 83,002 SNPs and 9,169 diseases and phenotypes (See TABLE 3).

3. METHODS

3.1. Filtering scheme for extraction of rare -damaging variants

3.1.1. Analysis focused on variants

We focused on rare and deleterious variants, and therefore, it was necessary to distinguish variants that have very low frequencies and deleterious predictive scores (See FIGURE 1).

Rare variants were determined by known frequencies through the 1000 Genomes Project. Based on the allele frequencies of the 1000 Genomes Project, variants with minor allele frequencies less than 1% were obtained.

These rare variants were filtered using three predictive scores indicating the degree of deleteriousness of the variants. The scores used were SIFT, PolyPhen-2, and CADD.

SIFT(Sorting Intolerant From Tolerant) is a predictive score based on the assumption that sequence changes in proteins due to evolution would be related to the structure and function of the protein. In other words, if a particular sequence is important for the function of the protein, it should be conserved, and if not, it would show various changes[20]. SIFT is divided into Deleterious(SIFT <= 0.05) and Tolerated(SIFT > 0.05) depending on the score. Therefore, variants with SIFT score of less than or equal to 0.05 were selected.

PolyPhen-2(Polymorphism Phenotyping v2), like SIFT, is a tool that predicts possible impact of changes in protein function due to genetic variation, HumVar, one of the two datasets used to train and test PolyPhen-2 prediction models, was used in this study. It consisted of human disease-causing variants in UniProtKB and common human nsSNPs with MAF > 1% with no disease-associated annotation[21]. PolyPhen-2 is divided into three stages: Probably damaging(>= 0.909), Possibly damaging(0.447 <= pp2_hvar <= 0.909), and Benign(pp2_hvar <= 0.446). Variants with a score of less than 0.909 were removed.

CADD(Combined Annotation Dependent Depletion) has been developed as a tool for scoring the deleteriousness of single nucleotide variants in the human genome[22]. Variants with a score of more than or equal to 20 were obtained.

3.1.2. Analysis focused on gene

There may be one or more of variants in a gene. If more than one variant is present in a gene, the variants contained therein will include a deleterious variant. If two or more variants are present in one gene, it is possible that a deleterious or benign variant may be included. As more damaging variants are involved, the more dysfunctional a gene will be. Therefore, all variants with SIFT score annotation were used to select genes that are expected to be more malfunctioning. And the intersection of two results, analysis focused on variants and analysis focused on gene, was obtained. In other words, rare-damaging variants in the gene that are expected to be abnormal were obtained (See FIGURE 1).

3.2. Annotation of information

3.2.1. Annotation with ANNOVAR

In the annotation process using ANNOVAR, three databases were used: dbnsfp30a, which contains information about the scores indicating the degeree of deleteriousness of variants, knownGene, which contains information about the gene containing the variant, and 1000g2015aug_all, which is used for annotation to allele frequencies (See TABLE 1).

3.2.2. Annotation with Oncotator

Oncotator was used to perform additional annotation on rare-damaging variants present in genes that are expected to have functional abnormalities obtained through filtering. While annotation was achievable directly from publicly available resources, Oncotator was used for quicker and easier annotation. Oncotator has been developed to assist in the analysis of cancer-related genes, but since it has been confirmed to work in other SNPs, it could also be used as an annotation tool for RarePedia (See TABLE 2).

3.3. Generation of HTML pages about rare -damaging variants

All information about the resultant variants was written in HTML format. One HTML page was created for one variant, and in this process, additional information obtained through the web parsing was added (See TABLE 3).

3.4. Validation with publicly available Next Generation Sequencing(NGS) data

3.4.1. Statistics from 1000 Genomes Project data

First, we identified how many rare-damaging variants were found in all of the 1000 Genomes Project sequencing data, and how many rare-damaging variants there were for 2,504 individuals, on average.

For obtaining rare variants, variants with allele frequencies of less than 1% were extracted. And damaging variants with SIFT score of 0.05 or less, PolyPhen-2 HVAR score of 0.909 or greater, and CADD score of 20 or greater were obtained.

3.4.2. Statistics from Alzheimer's Disease Sequencing Project(ADSP) data

Validation using next generation sequencing data of 651 EOAD patients was performed. Like validation using 1000 Genomes Project sequencing data, rare-damaging variants were obtained through the same filtering conditions using Alzheimer's Disease Sequencing Project(ADSP) data.

4. RESULT

4.1. Statistics about rare-damaging variants from Next Generation Sequencing Data

4.1.1. Statistics from 1000 Genomes Project data

70,281,123 rare variants were found in 1000 Genomes Project data.. while the total number of damaging variants obtained through annotation and filtering was 425. A total of 85 rare-damaging variants are the intersection of these two results. There was only one sample that had all 85 rare-damaging variants, and 496 samples that had no rare-damaging variant. It has also been confirmed that most of the samples had one rare-damaging variant. The average was 1.13738(±1.882775) (See FIGURE 2).

The result obtained do not include variants with an allele frequency of 0. However, when sequencing genomes of individuals, variants that are absent in the 1000 Genomes Project can be bound.

4.1.2. Statistics from Alzheimer's Disease Sequencing Project(ADSP) data

From the data of early-onset Alzheimer's Disease(EOAD) patients, a total of 23,635 rare-damaging variants were identified, which indicates that, on average, there are about 36 rare-damaging variants in one sample, even if there is no overlap between samples. In practice, one sample has an average of 81.298(±10.31878) rare-damaging variants (See FIGURE 3a). Also, 15,297 damaging variants with allele frequencies of 0 were identified (See FIGURE 3b).

4.1.3. Additional Whole-Exome Sequencing(WXS) data

To verify the actual results of RarePedia, we performed validation using randomly selected 10 Korean whole exome sequencing(WXS) data.

There were a total of 988 rare-damaging variants obtained. 10 samples were found to have an average of 153.7(±13.27529003) rare-damaging variants (See FIGURE 4a). This result included 554 rare variants not found in 1000 Genomes Project data, with an average of 104.5(±11.77803983) per sample (See FIGURE 4b).

4.1.4. ClinVar information about variants

In addition, we confirmed the presence of ClinVar information on the 1000 Genomes Project data, Alzheimer's Disease Sequencing Project(ADSP) data, and additional 10 Korean whole-exome sequencing(WXS) data used for validation.

ClinVar provides the relationship between variants and phenotypes, particularly Mendelian disorders[23]. In ClinVar, clinical significance value is divided into 14 classes for known/unknown information (See

TABLE 4).

1000 Genomes Project data, ADSP data, and 10 Korean data show similar trends. Benign variants were the most common for known information, but the number of variants with no information was overwhelming(See FIGURE 5).

4.2. Information contained in the HTML page generated by RarePedia

RarePedia is diveded into six sections, which helps users to view large amounts of data in a single page. Starting with a rare-damaging variant obtained through filtering steps, it shows a series of information about the variant, gene, transcript, protein, pathway and phenotype(disease). The description of the six sections is given below (See TABLE 5).

Information about the rare-damaging variant obtained through RarePedia is contained in the “Variant Section” (See FIGURE 6a). Starting with the basic information contained in the VCF used as the input, information from annotation, such as various scores of the variant and allele frequency, are displayed.

The “Gene section” shows information about the gene containing the rare-damaging variant introduced in “Variant section”. Gene name and description, various gene functions, and literature related to the gene are introduced in this section (See FIGURE 6b).

The “Transcript Section” provides information on the transcript used as a measure of gene expression. It also includes information provided by Expression Atlas and The Human Protein Atlas, which show how many copies of the transcript are expressed in various body tissues. Furthermore, it is possible to confirm how large quantities of copies of the transcript are expressed in cancer tissues is shown through The Human Protein Atlas (See FIGURE 6c).

If the transcript has protein-coding sequence, it is translated in the amino acid sequence of a protein. The “Protein Section” provides names of the protein used in different databases, and size, family and domains of the protein (See FIGURE 6d).

A protein is involved in various pathways in the body together with many substances such as proteins, RNAs, and ions. One proteins is not involved in only one pathway, and pathway information is provided in several databases such as KEGG and Reactome. GeneCards shows the information provided by several pathway databases in one place and is used in RarePedia for this reason (See FIGURE 6e).

Lastly, the “Phenotype Section” provides information on diseases known to be associated with the gene. The database provided by DisGeNET is used, and the disease ID, disease name, ad source are provided (See FIGURE 6f).

5. DISCUSSION

Current version of RarePedia(v1.0) is close to a prototype. It is a tool that receives a VCF as input and provides information about all rare-damaging variants in the individual. For a single rare-damaging variant, a large amount of information consisting of six sections(Variant, Gene, Transcript, Protein, Pathway and Phenotype) is written as a single HTML page.

Oncotator is a very useful tool to annotate a lot of information simultaneously. Therefore, when the initial RarePedia was devised, we wanted to use Oncotator as a way to annotate at once. Oncotator annotate a large amount of information consisting of more than 200 columns for one variant, and the time required for 1000 variants was approximately 20 seconds. It would be very beneficial if there were a small number of variants, but it would have taken a very long time if there were many. Also, in RarePedia, variants that are not rare-damaging variants do not need to have much information. In other words, there was a limit in that it takes much time for annotation of excess information that would not be used. Therefore, we used a two-step approach, where we first annotated minimum information for filtering and added a large amount of information to the resulting rare-damaging variants using Oncotator.

Because ANNOVAR does not accept VCF as an input, the input VCF file needs to be converted to an input type used by ANNOVAR. If annotating the VCF with 600,000 SNPs using Oncotator would take more than three hours. However, as a result of the strategy being modified to use ANNOVAR first, it has been confirmed that the time required for the entire RarePedia process using the same data was reduced to about 30 minutes. In other words, the required time was shortened by more than six-fold.

In the results from 1000 Genomes Project data, most of the samples had zero or one rare-damaging variants, except one sample that had 85 rare-damaging variants. Because sequencing targets of 1000 Genomes Project are healthy adults, it would be of no surprise that there are few rare-damaging variants.

In general, Alzheimer's disease(AD) is considered to be closely related to aging[23]. Early-onset Alzheimer's disease(EOAD), which is thought to be related to genetic factors, makes up about 5% of total AD[24]. In the Alzheimer's Disease Sequencing Project(ADSP) data, 651 out of 5,777 AD patients were identified as EOAD patients (about 11%).

Analysis of 615 patient data showed that on average, approximately 80 rare-damaging variants were present. There were a total of 23,635 rare-damaging variants, of which 15,297 variants without allele frequencies obtained from 1000 Genomes Project were found.

As a result of analysis using 10 Korean sample data, it showed similar trend to EOAD data. An average of 150 rare-damaging variants were found in 10 Korean samples, about twice as many as EOAD data. Most of the data provided by ADSP were from Caucasians, and the EOAD patients data contained in it were also the same. We cannot be certain because the number of samples analyzed is small, but this could possibly be due to racial differences.

In addition, through ClinVar annotation, we confirmed that there are not many variants known to be associated with disease yet. Of the 85 rare-damaging variants obtained from the 1000 Genomes Project data, only five variants were annotated the information using the ClinVar database. In the ADSP data, 1,053 variants out of 23,635 were annotated, and in 10 Korean WXS data, 40 out of 988 were. Only about 5% of all rare-damaging variants have ClinVar information. Of course, it would be a very high percentage when compared with the whole data. However, it might mean that there is still less information about the large number of variants obtained through next generation sequencing. Therefore, the information of variants that are revealed in the future needs to be combined, and many research institutes are now doing such information integration.

Currently, RarePedia focuses on individual genomes. It is a useful tool to help users see large amount of known information, such as gene, transcript, protein, pathway, and phenotypes, associated with one rare-damaging variant in an individual's genome. The ultimate goal of RarePedia is to accumulate information related to rare-damaging variants found in a small number of individuals, and to act as a window to link these variants to phenotype such as diseases.

6. REFERENCE

1. Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., ... & Boix-Chornet, M. Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A. 2005 Jul 26;102(30):10604-9. Epub 2005 Jul 11.

2. Park, L., & Kim, J. H. Rare high-impact disease variants: properties and identifications. Genet Res(Camb). 2016 Mar 21;98:e6.

3. Lin, J. R., Zhang, Q., Cai, Y., Morrow, B. E., & Zhang, Z. D. Integrated rare variant-based risk gene prioritization in disease case-control sequencing studies. PLoS Genet. 2017 Dec 27;13(12):e1007142.

4. Siva, N. 1000 Genomes project. Nat Biotechnol. 2008 Mar;26(3):256.

5. Beecham, G. W., Bis, J. C., Martin, E. R., Choi, S. H., DeStefano, A. L., van Duijn, C. M., ... & Naj, A. C. The Alzheimer's Disease Sequencing Project: Study design and sample selection. Neurol Genet. 2017 Oct 13;3(5):e194.

6. Wang, K., Li, M., & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. BMC Bioinformatics. 2017 Jul 17;18(1):341.

7. Ramos, A. H., Lichtenstein, L., Gupta, M., Lawrence, M. S., Pugh, T. J., Saksena, G., ... & Getz, G. Oncotator: cancer variant annotation tool. Hum Mutat. 2015 Apr;36(4):E2423-9.

8. Canese, K., & Weis, S. PubMed: the bibliographic database. The NCBI Handbook. 2013 March 20

9. Canese K. PubMed Celebrates its 10th Anniversary! NLM Tech Bull. 2006 Sep-Oct;(352):e5.

10. Mitchell, J. A., Aronson, A. R., Mork, J. G., Folk, L. C., Humphrey, S. M., & Ward, J. M. Gene indexing: characterization and analysis of NLM's GeneRIFs. AMIA Annu Symp Proc. 2003; 2003: 460 – 464.

11. Papatheodorou, I., Fonseca, N. A., Keays, M., Tang, Y. A., Barrera, E., Bazant, W., ... & Huerta, L. Expression Atlas: gene and protein expression across multiple studies and organisms. Nucleic Acids Res. 2018 Jan 4; 46(Database issue): D246 – D251. Published online 2017 Nov 20.

12. Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., ... & Olsson, I. Tissue-based map of the human proteome. Science. 2015 Jan 23;347(6220):1260419.

13. Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., ... & Sanli, K. A pathology atlas of the human cancer transcriptome. Science. 2017 Aug 18;357(6352). pii: eaan2507.

14. UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017 Jan 4; 45(Database issue): D158–D169. Published online 2016 Nov 28.

15. Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., ... & Gough, J. InterPro in 2017—beyond protein family and domain annotations. Nucleic Acids Res. 2017 Jan 4; 45(Database issue): D190–D199. Published online 2016 Nov 28.

16. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., ... & Durbin, R. The Ensembl genome database project. Nucleic Acids Res. 2002 Jan 1; 30(1): 38–41.

17. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., & Lancet, D. GeneCards: integrating information about genes, proteins and diseases. GENETWORK. Volume 13, Issue 4, p163, April 1997

18. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., & Lancet, D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics. 1998;14(8):656-64.

19. Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., ... & Furlong, L. I. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database(Oxford). 2015; 2015: bav028. Published online 2015 Apr 15.

20. Ng, P. C., & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1; 31(13): 3812–3814.

21. Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., ... & Sunyaev, S. R. A method and server for predicting damaging missense mutations. Nat Methods. Author manuscript; available in PMC 2010 Oct 1. Published in final edited form as: Nat Methods. 2010 Apr; 7(4): 248–249.

22. Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., & Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. Author manuscript; available in PMC 2014 Sep 1. Published in final edited form as: Nat Genet. 2014 Mar; 46(3): 310–315. Published online 2014 Feb 2.

23. Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014 Jan 1; 42(Database issue): D980–D985. Published online 2013 Nov 14.

24. Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., ... & Snyder, P. J. The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer′s disease. Alzheimers Dement. Author manuscript; available in PMC 2012 Mar 25. Published in final edited form as: Alzheimers Dement. 2011 May; 7(3): 270–279. Published online 2011 Apr 21.

25. Zhu, X. C., Tan, L., Wang, H. F., Jiang, T., Cao, L., Wang, C., ... & Yu, J. T. Rate of early onset Alzheimer's disease: a systematic review and meta-analysis. Ann Transl Med. 2015 Mar; 3(3): 38.

TABLES

TABLE 1. Databases used by ANNOVAR

Build	Table Name	Explanation
hg19	dbnsfp30a	whole−exome SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, MetaSVM, MetaLR, VEST, CADD, GERP++, DANN, fitCons, PhyloP and SiPhy scores from dbNSFP version 3.0a
hg19	1000g2015aug_all	alternative allele frequency data in 1000 Genomes Project for autosomes(ALL, AFR(African), AMR(Admixed American), EAS(East Asian), EUR(European), SAS(South Asian)).
hg19	knownGene	FASTA sequences for all annotated transcripts in UCSC Known Gene
hg19	clinvar_20160302	Clinvar version 20160302 with separate columns (CLINSIG CLNDBN CLNACC CLNDSDB CLNDSDBID)

TABLE 2. Databases used by Oncotator

Index	MAF Column Header	Description of Values
35	Genome_Change	String describing '+' strand genomic coordinates and alleles.
36	Annotation_Transcript	Ensembl transcript ID of transcript used for annotation.
37	Transcript_Strand	Strand orientation of the above transcript.
38	Transcript_Exon	Indicates the exon affected by the mutation.
39	Transcript_Position	Describes absolute start and end coordinates(separated by a underscore characer) with respect to the reference transcript used in the Annotation_Transcript column. Note these coordinates will differ from the coding region coordinates used in the cDNA_Change and Codon_Change columns. Only one number will be provided if the start and end coordinates are the same.
40	cDNA_Change	Coding positon and alleles. Coordinates are coding sequence coordinates.
41	Codon_Change	String describing transcript coordinates and alleles in context of codon sequences involved.
42	Protein_Change	Protein postion and alleles involved.
43	Other_Transcripts	HUGO symbol, Ensembl transcript id, variant classifcation and protein change of other transcripts overlapping with mutation.
44	Refseq_mRNA_Id	RefSeq transcript ID.
45	Refseq_prot_Id	Refseq protein ID.
46	SwissProt_acc_Id	UniProt accession ID.
47	SwissProt_entry_Id	UniProt entry name ID.
48	Description	If available, description text for transcript.
49	UniProt_AApos	UniProt protein position used to derive position-specific annotations. This can differ from the protein position listed in the 'Protein_Change' field if the UCSC and Uniprot protein sequeneces differ.
50	UniProt_Region	Overlapping UniProt regions of interest(e.g. functional domain or repeat region).
51	UniProt_Site	Overlapping UniProt single amino acid sites of interest(e.g. cleavage or inhibitory sites for proteases).
52	UniProt_Natural_Variations	Overlapping UniProt variants of interest(e.g. polymorphisms or disease-associated mutations).
53	UniProt_Experimental_Info	Overlapping UniProt sites with experimental data(e.g. mutagenesis data leading to protein activity inhibition).
54	GO_Biological_Process	Gene Ontology terms describing pathways and processes UniProt protein is involved in.
55	GO_Cellular_Component	Gene Ontology terms describing localization of given UniProt protein.
56	GO_Molecular_Function	Gene Ontology terms describing molecular activity of given UniProt protein.
57	COSMIC_overlapping_mutation	Protein changes of overlapping alterations.

	s	Number of samples in COSMIC with said mutation is in parentheses.
58	COSMIC_fusion_genes	Gene symbols of fusion events involving gene in COSMIC. Number of samples in COSMIC with said mutation is in parentheses.
59	COSMIC_tissue_types_affected	Tissue type summary of tumor samples involving gene in COSMIC. Number of samples in COSMIC is in parentheses.
60	COSMIC_total_alterations_in_gene	Total numbers of records for gene in COSMIC
61	Tumorscape_Amplification_Peaks	Overlapping significant GISTIC aplification focal peaks from Tumorscape.(Number of genes in peak and q-value of peaks is in parentheses). Only peak regions with a q-value <= 0.20 are reported.
62	Tumorscape_Deletion_Peaks	Overlapping significant GISTIC deletion focal peaks from Tumorscape.(Number of genes in peak and q-value of peaks is in parentheses). Only peak regions with a q-value <= 0.20 are reported.
63	TCGAscape_Amplification_Peaks	Overlapping significant GISTIC amplification focal peaks from TCGAscape(Number of genes in peak and q-value of peaks is in parentheses). Only peak regions with a q-value <= 0.20 are reported.
64	TCGAscape_Deletion_Peaks	Overlapping significant GISTIC deletion focal peaks from TCGAscape.(Number of genes in peak and q-value of peaks is in parentheses). Only peak regions with a q-value <= 0.20 are reported.
65	DrugBank	Listing of compounds from DrugBank known to interact with genes(DrugBank compound ID in parentheses).
66	ref_context	Genomic sequence at variant locus with additional 10 bp of flanking sequence on either side.
67	gc_content	Fraction of G or C bases in flanking 100 bp of variant locus
68	CCLE_ONCOMAP_overlapping_mutations	Protein change of overlapping mutations in CCLE Oncomap dataset. Cell line name and lineage are provided in parentheses.
69	CCLE_ONCOMAP_total_mutations_in_gene	Total number of mutations in CCLE Oncomap data for this gene.
70	CGC_Mutation_Type	
71	CGC_Translocation_Partner	Known translocation partner gene as reported in Cancer Gene Census
72	CGC_Tumor_Types_Somatic	
73	CGC_Tumor_Types_Germline	
74	CGC_Other_Diseases	Other diseases/syndromes with alterations in this gene as reported in Cancer Gene Census.
75	DNARepairGenes_Role	
76	FamilialCancerDatabase_Syndromes	
77	MUTSIG_Published_Results	Published MutSig analyses with gene in signifcant results. Gene rank and q-value are provided in parentheses.
78	OREGANNO_ID	ID for ORegAnno regulatory regions,

		transcription factor binding sites, and regulatory polymorphisms as reported in the UCSC Genome Browser.
79	OREGANNO_Values	
80	1000Genome_AA	
81	1000Genome_AC	1000 Genomes annotation, Alternate Allele Count
82	1000Genome_AF	1000 Genomes annotation, Global Allele Frequency based on AC/AN"
83	1000Genome_AFR_AF	1000 Genomes annotation, Allele Frequency for samples from AFR based on AC/AN
84	1000Genome_AMR_AF	1000 Genomes annotation, Allele Frequency for samples from AMR based on AC/AN
85	1000Genome_AN	1000 Genomes annotation, Total Allele Count
86	1000Genome_ASN_AF	1000 Genomes annotation, Allele Frequency for samples from ASN based on AC/AN
87	1000Genome_AVGPOST	1000 Genomes annotation, Average posterior probability from MaCH/Thunder
88	1000Genome_CIEND	1000 Genomes annotation, Confidence interval around END for imprecise variants
89	1000Genome_CIPOS	1000 Genomes annotation, Confidence interval around POS for imprecise variants
90	1000Genome_END	1000 Genomes annotation, End position of the variant described in this record
91	1000Genome_ERATE	1000 Genomes annotation, Per-marker Mutation rate from MaCH/Thunder
92	1000Genome_EUR_AF	1000 Genomes annotation, Allele Frequency for samples from EUR based on AC/AN
93	1000Genome_HOMLEN	1000 Genomes annotation, Length of base pair identical micro-homology at event breakpoints
94	1000Genome_HOMSEQ	1000 Genomes annotation, Sequence of base pair identical micro-homology at event breakpoints
95	1000Genome_LDAF	1000 Genomes annotation, MLE Allele Frequency Accounting for LD
96	1000Genome_RSQ	1000 Genomes annotation, Genotype imputation quality from MaCH/Thunder
97	1000Genome_SNPSOURCE	1000 Genomes annotation, indicates if a snp was called when analysing the low coverage or exome alignment data
98	1000Genome_SVLEN	1000 Genomes annotation, Difference in length between REF and ALT alleles
99	1000Genome_SVTYPE	1000 Genomes annotation, Type of structural variant
100	1000Genome_THETA	1000 Genomes annotation, Per-marker Transition rate from MaCH/Thunder
101	1000Genome_VT	1000 Genomes annotation, indicates what type of variant the line represents
102	ACHILLES_Lineage_Results_Top_Genes	Lineages in ACHILLES dataset with gene in top 200 scoring genes. Rank score of gene followed by individual hairpin ranks for given gene are provided in parentheses.
103	CGC_Cancer Germline Mut	Cancer Gene Census annotation, "yes" if variant is in a gene that is mutated in the germline predisposing to cancer.

104	CGC_Cancer Molecular Genetics	Cancer Gene Census annotation, Indicates whether variants in mutated gene are dominant or recessive.
105	CGC_Cancer Somatic Mut	Cancer Gene Census annotation, "yes" if variant is in a gene that is somatically mutated in cancer.
106	CGC_Cancer Syndrome	Cancer related syndromes with alterations in this gene as reported in Cancer Gene Census.
107	CGC_Chr	Cancer Gene Census annotation, Chromosome.
108	CGC_Chr Band	Cancer Gene Census annotation, Chromosome band.
109	CGC_GeneID	Cancer Gene Census annotation, Entrez gene ID.
110	CGC_Name	Cancer Gene Census annotation, Full gene name.
111	CGC_Other Germline Mut	Cancer Gene Census annotation, "yes" if variant is in a gene that is germline mutated in other diseases/syndromes.
112	CGC_Tissue Type	Cancer Gene Census annotation, Tissue types with mutations in this gene.
113	COSMIC_n_overlapping_mutations	Total number of COSMIC mutations at variant site.
114	COSMIC_overlapping_mutation_descriptions	COSMIC mutation descriptions at variant site. Number of samples in COSMIC is in parentheses.
115	COSMIC_overlapping_primary_sites	Primary site summary of tumor samples with COSMIC mutations at variant site. Number of samples in COSMIC is in parentheses.
116	ClinVar_ASSEMBLY	ClinVar annotation, Assembly
117	ClinVar_HGMD_ID	ClinVar annotation, HGNMD ID
118	ClinVar_SYM	ClinVar annotation, Gene symbol
119	ClinVar_TYPE	ClinVar annotation, Type
120	ClinVar_rs	ClinVar annotation, dbSNP ID
121	ESP_AA	ESP annotation, chimpAllele
122	ESP_AAC	ESP annotation, aminoAcidChange
123	ESP_AA_AC	ESP annotation, African American Allele Count in the order of AltAlleles,RefAllele. For INDELs, A1, A2, or An refers to the N-th alternate allele while R refers to the reference allele.
124	ESP_AA_AGE	ESP annotation, Estimated Variant Age in kilo years for the African American Population
125	ESP_AA_GTC	ESP annotation, African American Genotype Counts in the order of listed GTS
126	ESP_AvgAAsampleReadDepth	ESP annotation, Mean read depth at variant position in African American ESP cohort.
127	ESP_AvgEAsampleReadDepth	ESP annotation, Mean read depth at variant position in European American ESP cohort.
128	ESP_AvgSampleReadDepth	ESP annotation, Mean read depth at variant position in all ESP samples.
129	ESP_CA	ESP annotation, clinicalAssociation
130	ESP_CDP	ESP annotation, cDNAPosition
131	ESP_CG	ESP annotation, consScoreGERP
132	ESP_CP	ESP annotation, scorePhastCons

133	ESP_Chromosome	ESP annotation, Chromosome
134	ESP_DBSNP	ESP annotation, dbSNP version which established the rs_id
135	ESP_DP	ESP annotation, Average Sample Read Depth"
136	ESP_EA_AC	ESP annotation, European American Allele Count in the order of AltAlleles,RefAllele. For INDELs, A1, A2, or An refers to the N-th alternate allele while R refers to the reference allele.
137	ESP_EA_AGE	ESP annotation, Esitmated Variant Age in kilo years for the European American Population
138	ESP_EA_GTC	ESP annotation, European American Genotype Counts in the order of listed GTS
139	ESP_EXOME_CHIP	ESP annotation, Whether a SNP is on the Illumina HumanExome Chip
140	ESP_FG	ESP annotation, functionGVS
141	ESP_GL	ESP annotation, geneList
142	ESP_GM	ESP annotation, accession
143	ESP_GS	ESP annotation, granthamScore
144	ESP_GTC	ESP annotation, Total Genotype Counts in the order of listed GTS
145	ESP_GTS	ESP annotation, Observed Genotypes. For INDELs, A1, A2, or An refers to the N-th alternate allele while R refers to the reference allele.
146	ESP_GWAS_PUBMED	ESP annotation, PubMed records for GWAS hits
147	ESP_MAF	ESP annotation, Minor Allele Frequency in percent in the order of EA,AA,All
148	ESP_PH	ESP annotation, polyPhen
149	ESP_PP	ESP annotation, proteinPosition"
150	ESP_Position	ESP annotation, Genomic position"
151	ESP_TAC	ESP annotation, Total Allele Count in the order of AltAlleles,RefAllele For INDELs, A1, A2, or An refers to the N-th alternate allele while R refers to the reference allele.
152	ESP_TotalAAsamplesCovered	ESP annotation, Total African American samples with read coverage at variant site.
153	ESP_TotalEAsamplesCovered	ESP annotation, Total European American samples with read coverage at variant site.
154	ESP_TotalSamplesCovered	ESP annotation, Total ESP samples with read coverage at variant site.
155	Ensembl_so_accession	Ensembl Sequence ontology accession
156	Ensembl_so_term	Ensembl Sequence ontology term
157	Familial_Cancer_Genes_Reference	Familial cancer database reference used.
158	Familial_Cancer_Genes_Synonym	
159	HGNC_Ensembl Gene ID	
160	HGNC_HGNC ID	
161	HGNC_RefSeq IDs	
162	HGNC_Status	

163	HGNC_UCSC ID(supplied by UCSC)	
164	HGVS_coding_DNA_change	HGVS compliant string describing coding positon and alleles.
165	HGVS_genomic_change	HGVS compliant string describing '+' strand genomic coordinates and alleles.
166	HGVS_protein_change	HGVS compliant string describing protein postion and alleles involved.
167	ORegAnno_bin	UCSC Genome Browser bin for ORegAnno entry.
168	UniProt_alt_uniprot_accessions	Alternative UniProt accession IDs
169	build	User-supplied build value
170	ccds_id	
171	dbNSFP_1000Gp1_AC	dbNSFP annotation, Alternative allele counts in the whole 1000 genomes phase 1(1000Gp1) data.
172	dbNSFP_1000Gp1_AF	dbNSFP annotation, Alternative allele frequency in the whole 1000Gp1 data.
173	dbNSFP_1000Gp1_AFR_AC	dbNSFP annotation, Alternative allele counts in the 1000Gp1 African descendent samples.
174	dbNSFP_1000Gp1_AFR_AF	dbNSFP annotation, Alternative allele frequency in the 1000Gp1 African descendent samples.
175	dbNSFP_1000Gp1_AMR_AC	dbNSFP annotation, Alternative allele counts in the 1000Gp1 American descendent samples.
176	dbNSFP_1000Gp1_AMR_AF	dbNSFP annotation, Alternative allele frequency in the 1000Gp1 American descendent samples.
177	dbNSFP_1000Gp1_ASN_AC	dbNSFP annotation, Alternative allele counts in the 1000Gp1 Asian descendent samples.
178	dbNSFP_1000Gp1_ASN_AF	dbNSFP annotation, Alternative allele frequency in the 1000Gp1 Asian descendent samples.
179	dbNSFP_1000Gp1_EUR_AC	dbNSFP annotation, Alternative allele counts in the 1000Gp1 European descendent samples.
180	dbNSFP_1000Gp1_EUR_AF	dbNSFP annotation, Alternative allele frequency in the 1000Gp1 European descendent samples.
181	dbNSFP_Ancestral_allele	dbNSFP annotation, Ancestral allele(based on 1000 genomes reference data). The following comes from its original README file"
182	dbNSFP_CADD_phred	dbNSFP annotation, CADD phred-like score. This is phred-like rank score based on whole genome CADD raw scores. Please refer to Kircher et al.(2014) Nature Genetics 46(3)
183	dbNSFP_CADD_raw	dbNSFP annotation, CADD raw score for funtional prediction of a SNP. Please refer to Kircher et al.(2014) Nature Genetics 46(3)
184	dbNSFP_CADD_raw_rankscore	dbNSFP annotation, CADD raw scores were ranked among all CADD raw scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of CADD raw scores in dbNSFP. Please note the following copyright statement for CADD"

185	dbNSFP_ESP6500_AA_AF	dbNSFP annotation, Alternative allele frequency in the Afrian American samples of the NHLBI GO Exome Sequencing Project(ESP6500 data set).
186	dbNSFP_ESP6500_EA_AF	dbNSFP annotation, Alternative allele frequency in the European American samples of the NHLBI GO Exome Sequencing Project(ESP6500 data set).
187	dbNSFP_Ensembl_geneid	dbNSFP annotation, Ensembl gene id"
188	dbNSFP_Ensembl_transcriptid	dbNSFP annotation, Ensembl transcript ids(separated by ";")
189	dbNSFP_FATHMM_pred	dbNSFP annotation, If a FATHMMori score is <=-1.5(or rankscore <=0.81415) the corresponding NS is predicted as "D(AMAGING)"; otherwise it is predicted as "T(OLERATED)". Multiple predictions separated by ";"
190	dbNSFP_FATHMM_rankscore	dbNSFP annotation, FATHMMori scores were ranked among all FATHMMori scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of FATHMMori scores in dbNSFP. If there are multiple scores, only the most damaging(largest) rankscore is presented. The scores range from 0 to 1.
191	dbNSFP_FATHMM_score	dbNSFP annotation, FATHMM default score(weighted for human inherited-disease mutations with Disease Ontology)(FATHMMori). Scores range from -18.09 to 11.0. Multiple scores separated by ";" Please refer to Shihab et al.(2013) Human Mutation 34(1)"
192	dbNSFP_GERP++_NR	dbNSFP annotation, GERP++ neutral rate"
193	dbNSFP_GERP++_RS	dbNSFP annotation, GERP++ RS score, the larger the score, the more conserved the site.
194	dbNSFP_GERP++_RS_rankscore	dbNSFP annotation, GERP++ RS scores were ranked among all GERP++ RS scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of GERP++ RS scores in dbNSFP.
195	dbNSFP_Interpro_domain	dbNSFP annotation, domain or conserved site on which the variant locates. Domain annotations come from Interpro database. The number in the brackets following a specific domain is the count of times Interpro assigns the variant position to that domain, typically coming from different predicting databases. Multiple entries separated by ";".
196	dbNSFP_LRT_Omega	dbNSFP annotation, estimated nonsynonymous-to-synonymous-rate ratio(Omega, reported by LRT)"
197	dbNSFP_LRT_converted_rankscore	dbNSFP annotation, LRTori scores were first converted as LRTnew=1-LRTori*0.5 if Omega<1, or LRTnew=LRTori*0.5 if Omega>=1. Then LRTnew scores were ranked among all LRTnew scores in dbNSFP. The rankscore is the ratio of the rank over the total number of the scores in dbNSFP. The scores range from 0.00166 to 0.85682.
198	dbNSFP_LRT_pred	dbNSFP annotation, LRT prediction,

		D(eleterious), N(eutral) or U(nknown), which is not solely determined by the score. "
199	dbNSFP_LRT_score	dbNSFP annotation, The original LRT two-sided p-value(LRTori), ranges from 0 to 1.
200	dbNSFP_LR_pred	dbNSFP annotation, Prediction of our LR based ensemble prediction score,"T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0.5. The rankscore cutoff between "D" and "T" is 0.82268.
201	dbNSFP_LR_rankscore	dbNSFP annotation, LR scores were ranked among all LR scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of LR scores in dbNSFP. The scores range from 0 to 1.
202	dbNSFP_LR_score	dbNSFP annotation, Our logistic regression(LR) based ensemble prediction score, which incorporated 10 scores(SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 genomes populations. Larger value means the SNV is more likely to be damaging. Scores range from 0 to 1.
203	dbNSFP_MutationAssessor_pred	dbNSFP annotation, MutationAssessor's functional impact of a variant "
204	dbNSFP_MutationAssessor_rankscore	dbNSFP annotation, MAori scores were ranked among all MAori scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of MAori scores in dbNSFP. The scores range from 0 to 1.
205	dbNSFP_MutationAssessor_score	dbNSFP annotation, MutationAssessor functional impact combined score(MAori). The score ranges from -5.545 to 5.975 in dbNSFP. Please refer to Reva et al.(2011) Nucl. Acids Res. 39(17)"
206	dbNSFP_MutationTaster_converted_rankscore	dbNSFP annotation, The MTori scores were first converted"
207	dbNSFP_MutationTaster_pred	dbNSFP annotation, MutationTaster prediction, "A"("disease_causing_automatic"), "D"("disease_causing"), "N"("polymorphism") or "P"("polymorphism_automatic"). The score cutoff between "D" and "N" is 0.5 for MTori and 0.328 for the rankscore.
208	dbNSFP_MutationTaster_score	dbNSFP annotation, MutationTaster p-value(MTori), ranges from 0 to 1.
209	dbNSFP_Polyphen2_HDIV_pred	dbNSFP annotation, Polyphen2 prediction based on HumDiv, "D"("porobably damaging", HDIV score in[0.957,1] or rankscore in[0.52996,0.89917]), "P"("possibly damaging", HDIV score in[0.453,0.956] or rankscore in[0.34412,0.52842]) and "B"("benign", HDIV score in[0,0.452] or rankscore in[0.02656,0.34399]). Score cutoff for binary classification is 0.5 for HDIV score or 0.35411 for rankscore, i.e. the prediction is "neutral" if the HDIV score is smaller than 0.5(rankscore is smaller than 0.35411), and "deleterious" if the HDIV score is larger than

		0.5(rankscore is larger than 0.35411). Multiple entries are separated by ";".
210	dbNSFP_Polyphen2_HDIV_rankscore	dbNSFP annotation, Polyphen2 HDIV scores were first ranked among all HDIV scores in dbNSFP. The rankscore is the ratio of the rank the score over the total number of the scores in dbNSFP. If there are multiple scores, only the most damaging(largest) rankscore is presented. The scores range from 0.02656 to 0.89917.
211	dbNSFP_Polyphen2_HDIV_score	dbNSFP annotation, Polyphen2 score based on HumDiv, i.e. hdiv_prob. The score ranges from 0 to 1. Multiple entries separated by ";".
212	dbNSFP_Polyphen2_HVAR_pred	dbNSFP annotation, Polyphen2 prediction based on HumVar, "D"("probably damaging", HVAR score in[0.909,1] or rankscore in[0.62955,0.9711]), "P"("possibly damaging", HVAR in[0.447,0.908] or rankscore in[0.44359,0.62885]) and "B"("benign", HVAR score in[0,0.446] or rankscore in[0.01281,0.44315]). Score cutoff for binary classification is 0.5 for HVAR score or 0.45998 for rankscore, i.e. the prediction is "neutral" if the HVAR score is smaller than 0.5(rankscore is smaller than 0.45998), and "deleterious" if the HVAR score is larger than 0.5(rankscore is larger than 0.45998). Multiple entries are separated by ";".
213	dbNSFP_Polyphen2_HVAR_rankscore	dbNSFP annotation, Polyphen2 HVAR scores were first ranked among all HVAR scores in dbNSFP. The rankscore is the ratio of the rank the score over the total number of the scores in dbNSFP. If there are multiple scores, only the most damaging(largest) rankscore is presented. The scores range from 0.01281 to 0.9711.
214	dbNSFP_Polyphen2_HVAR_score	dbNSFP annotation, Polyphen2 score based on HumVar, i.e. hvar_prob. The score ranges from 0 to 1. Multiple entries separated by ";".
215	dbNSFP_RadialSVM_pred	dbNSFP annotation, Prediction of our SVM based ensemble prediction score,"T(olerated)" or "D(amaging)". The score cutoff between "D" and "T" is 0. The rankscore cutoff between "D" and "T" is 0.83357.
216	dbNSFP_RadialSVM_rankscore	dbNSFP annotation, RadialSVM scores were ranked among all RadialSVM scores in dbNSFP. The rankscore is the ratio of the rank of the screo over the total number of RadialSVM scores in dbNSFP. The scores range from 0 to 1.
217	dbNSFP_RadialSVM_score	dbNSFP annotation, Our support vector machine(SVM) based ensemble prediction score, which incorporated 10 scores(SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 genomes populations. Larger value means the SNV is more likely to be damaging. Scores range from -2 to 3 in dbNSFP.
218	dbNSFP_Reliability_index	dbNSFP annotation, Number of observed

		component scores(except the maximum frequency in the 1000 genomes populations) for RadialSVM and LR. Ranges from 1 to 10. As RadialSVM and LR scores are calculated based on imputed data, the less missing component scores, the higher the reliability of the scores and predictions.
219	dbNSFP_SIFT_converted_rankscore	dbNSFP annotation, SIFTori scores were first converted to SIFTnew=1-SIFTori, then ranked among all SIFTnew scores in dbNSFP. The rankscore is the ratio of the rank the SIFTnew score over the total number of SIFTnew scores in dbNSFP. If there are multiple scores, only the most damaging(largest) rankscore is presented. The rankscores range from 0.02654 to 0.87932.
220	dbNSFP_SIFT_pred	dbNSFP annotation, If SIFTori is smaller than 0.05(rankscore>0.55) the corresponding NS is predicted as "D(amaging)"; otherwise it is predicted as "T(olerated)". Multiple predictions separated by ";"
221	dbNSFP_SIFT_score	dbNSFP annotation, SIFT score(SIFTori). Scores range from 0 to 1. The smaller the score the more likely the SNP has damaging effect. Multiple scores separated by ";".
222	dbNSFP_SLR_test_statistic	dbNSFP annotation, SLR test statistic for testing natural selection on codons. A negative value indicates negative selection, and a positive value indicates positive selection. Larger magnitude of the value suggests stronger evidence.
223	dbNSFP_SiPhy_29way_logOdds	dbNSFP annotation, SiPhy score based on 29 mammals genomes. The larger the score, the more conserved the site.
224	dbNSFP_SiPhy_29way_logOdds_rankscore	dbNSFP annotation, SiPhy_29way_logOdds scores were ranked among all SiPhy_29way_logOdds scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of SiPhy_29way_logOdds scores in dbNSFP.
225	dbNSFP_SiPhy_29way_pi	dbNSFP annotation, The estimated stationary distribution of A, C, G and T at the site, using SiPhy algorithm based on 29 mammals genomes. "
226	dbNSFP_UniSNP_ids	dbNSFP annotation, rs numbers from UniSNP, which is a cleaned version of dbSNP build 129, in format"
227	dbNSFP_Uniprot_aapos	dbNSFP annotation, amino acid position as to Uniprot. Multiple entries separated by ";".
228	dbNSFP_Uniprot_acc	dbNSFP annotation, Uniprot accession number. Multiple entries separated by ";".
229	dbNSFP_Uniprot_id	dbNSFP annotation, Uniprot ID number. Multiple entries separated by ";".
230	dbNSFP_aaalt	dbNSFP annotation, alternative amino acid "." if the variant is a splicing site SNP(2bp on each end of an intron)
231	dbNSFP_aapos	dbNSFP annotation, amino acid position as to the protein. "-1" if the variant is a splicing site SNP(2bp on each end of an intron)

232	dbNSFP_aapos_FATHMM	dbNSFP annotation, ENSP id and amino acid positions corresponding to FATHMM scores. Multiple entries separated by ";"
233	dbNSFP_aapos_SIFT	dbNSFP annotation, ENSP id and amino acid positions corresponding to SIFT scores. Multiple entries separated by ";"
234	dbNSFP_aaref	dbNSFP annotation, reference amino acid. "." if the variant is a splicing site SNP(2bp on each end of an intron)
235	dbNSFP_cds_strand	dbNSFP annotation, coding sequence(CDS) strand(+ or −)
236	dbNSFP_codonpos	dbNSFP annotation, position on the codon(1, 2 or 3)
237	dbNSFP_fold−degenerate	dbNSFP annotation, degenerate type(0, 2 or 3)
238	dbNSFP_genename	dbNSFP annotation, gene name; if the NScan be assigned to multiple genes, gene names are separated by ";"
239	dbNSFP_hg18_pos(1−coor)	dbNSFP annotation, physical position on the chromosome as to hg18(1−based coordinate)
240	dbNSFP_phastCons100way_vertebrate	dbNSFP annotation, phastCons conservation score based on the multiple alignments of 100 vertebrate genomes(including human). The larger the score, the more conserved the site.
241	dbNSFP_phastCons100way_vertebrate_rankscore	dbNSFP annotation, phastCons100way_vertebrate scores were ranked among all phastCons100way_vertebrate scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phastCons100way_vertebrate scores in dbNSFP.
242	dbNSFP_phastCons46way_placental	dbNSFP annotation, phastCons conservation score based on the multiple alignments of 33 placental mammal genomes(including human). The larger the score, the more conserved the site.
243	dbNSFP_phastCons46way_placental_rankscore	dbNSFP annotation, phastCons46way_placental scores were ranked among all phastCons46way_placental scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phastCons46way_placental scores in dbNSFP.
244	dbNSFP_phastCons46way_primate	dbNSFP annotation, phastCons conservation score based on the multiple alignments of 10 primate genomes(including human). The larger the score, the more conserved the site.
245	dbNSFP_phastCons46way_primate_rankscore	dbNSFP annotation, phastCons46way_primate scores were ranked among all phastCons46way_primate scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phastCons46way_primate scores in dbNSFP.
246	dbNSFP_phyloP100way_vertebrate	dbNSFP annotation, phyloP(phylogenetic p−values) conservation score based on the multiple alignments of 100 vertebrate genomes(including human). The larger the score, the more conserved the site.
247	dbNSFP_phyloP100way_vertebrate_rankscore	dbNSFP annotation, phyloP100way_vertebrate scores were ranked among all

		phyloP100way_vertebrate scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phyloP100way_vertebrate scores in dbNSFP.
248	dbNSFP_phyloP46way_placental	dbNSFP annotation, phyloP(phylogenetic p-values) conservation score based on the multiple alignments of 33 placental mammal genomes(including human). The larger the score, the more conserved the site.
249	dbNSFP_phyloP46way_placental_rankscore	dbNSFP annotation, phyloP46way_placental scores were ranked among all phyloP46way_placental scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phyloP46way_placental scores in dbNSFP.
250	dbNSFP_phyloP46way_primate	dbNSFP annotation, phyloP(phylogenetic p-values) conservation score based on the multiple alignments of 10 primate genomes(including human). The larger the score, the more conserved the site.
251	dbNSFP_phyloP46way_primate_rankscore	dbNSFP annotation, phyloP46way_primate scores were ranked among all phyloP46way_primate scores in dbNSFP. The rankscore is the ratio of the rank of the score over the total number of phyloP46way_primate scores in dbNSFP.
252	dbNSFP_refcodon	dbNSFP annotation, reference codon
253	gencode_transcript_name	Gencode transcript name
254	gencode_transcript_status	Gencode transcript status
255	gencode_transcript_tags	Gencode transcript tags
256	gencode_transcript_type	Gencode transcript type
257	gene_id	Internal gene ID
258	gene_type	Type of gene used for variant annotation
259	havana_transcript	
260	secondary_variant_classification	Oncotator secondary variant classification
261	strand	Strand orientation of variant genomic coordinates
262	transcript_id	Transcript ID used for variant annotation

TABLE 3. Publicly available web-based Resources

Section	Resource	url
Gene	PubMed	https://www.ncbi.nlm.nih.gov/pubmed/
Gene	GeneRIF	https://www.ncbi.nlm.nih.gov/gene/about-generif
Transcript	Expression Atlas	https://www.ebi.ac.uk/gxa/home
Transcript	The Human Protein Atlas	https://www.proteinatlas.org/
Protein	UniProt	http://www.uniprot.org/
Protein	InterPro	http://www.ebi.ac.uk/interpro/
Protein	Ensembl GRCh37	http://grch37.ensembl.org/index.html
Pathway	GeneCards	www.genecards.org/
Phenotype	DisGeNET	www.disgenet.org/

TABLE 4. Options for Clinical significance on ClinVar submissions (SCV)

Clinical significance value	**Guidance for use in ClinVar SCV records**
Benign	As recommended by ACMG/AMP for variants interpreted for Mendelian disorders.
Likely benign	As recommended by ACMG/AMP for variants interpreted for Mendelian disorders.
Uncertain significance	As recommended by ACMG/AMP for variants interpreted for Mendelian disorders.
Likely pathogenic	As recommended by ACMG/AMP for variants interpreted for Mendelian disorders.
Pathogenic	As recommended by ACMG/AMP for variants interpreted for Mendelian disorders. Variants that have low penetrance may be submitted as "Pathogenic"; please also include information about the penetrance in a "Comment on clinical significance".
drug response	A general term for a variant that affects a drug response, not a disease. We anticipate adding more specific drug response terms based on a recommendation by CPIC.
association	For variants identified in a GWAS study and further interpreted for their clinical significance.
risk factor	For variants that are interpreted not to cause a disorder but to increase the risk.
protective	For variants that decrease the risk of a disorder, including infections.
Affects	For variants that cause a non-disease phenotype, such as lactose intolerance.

conflicting data from submitters	Only for submissions from a consortium, where groups within the consortium have conflicting intepretations of a variant but provide a single submission to ClinVar.
other	If ClinVar does not have the appropriate term for your submission, we ask that you submit "other" as clinical significance and contact us to discuss if there are other terms we should add.
not provided	For submissions without an interpretation of clinical significance. The primary goal of ClinVar is to archive reports of clinical significance of variants. Therefore submissions with a clinical significance of "not provided" should be limited to: "literature only" submissions that report a publication about the variant, without interpreting the clinical significance "research" submissions that provide functional significance (e.g. undetectable protein level) but no interpretation of clinical significance "clinical testing" or "phenotyping only" submissions from clinics or physicians that provide additional information about individuals with the variant, such as observed phenotypes, but do not interpret the clinical significance
Null	This value may not be submitted.

TABLE 5. Information provided by RarePedia

a. Variant Section

Information	Origin
Chromosome	input VCF
Position	input VCF
rsID	input VCF or information annotated by Oncotator
Reference Allele	input VCF
Alternative Allele	input VCF
Genotype	input VCF
Read Depth	input VCF
Quality	input VCF
SIFT score	information annotated by ANNOVAR
Polyphen2 HVAR	information annotated by ANNOVAR
CADD score	information annotated by ANNOVAR
GERP++_RS	information annotated by ANNOVAR
Variant Type	information annotated by Oncotator
Exon Number	information annotated by Oncotator
cDNA change	information annotated by Oncotator
Codon change	information annotated by Oncotator
Protein change	information annotated by Oncotator
Allele Frequncy	information annotated by ANNOVAR or Oncotator

b. Gene Section

Information	Origin
Gene Name	information annotated by Oncotator
Gene Symbol	information annotated by ANNOVAR of Oncotator
Synonyms	information annotated by Oncotator
ENSG	information annotated by Oncotator
Gene Location	Home - Gene - NCBI (https://www.ncbi.nlm.nih.gov/gene/)
Gene Type	information annotated by Oncotator
Gene Description	Home - Gene - NCBI (https://www.ncbi.nlm.nih.gov/gene/)
Gene Function	information annotated by Oncotator
Related articles in PubMed	Home - Gene - NCBI (https://www.ncbi.nlm.nih.gov/gene/)
GeneRIFs: Gene References Into Functions	Home - Gene - NCBI (https://www.ncbi.nlm.nih.gov/gene/)

c. Transcript Section

Information	Origin
Ensembl transcript ID	information annotated by Oncotator
Transcript position	information annotated by Oncotator
Transcript Strand	information annotated by ANNOVAR or Oncotator
Refseq mRNA ID	information annotated by Oncotator
Baseline expression	Expression Atlas (https://www.ebi.ac.uk/gxa/home)
Expression in other tissues	The Human Protein Atlas (https://www.proteinatlas.org/)
Expression in cancer tissues	The Human Protein Atlas (https://www.proteinatlas.org/)

d. Protein Section

Information	Origin
Recommeneded name	information annotated by Oncotator
UniProtKB/Swiss-Prot	information annotated by Oncotator
Ensembl proteins	information annotated by Oncotator
RefSeq protein	information annotated by Oncotator
Protein size	UniProt (https://www.uniprot.org/)
Protein family	InterPro (http://www.ebi.ac.uk/interpro/)
Protein damain	Ensembl GRCh37 (http://grch37.ensembl.org/index.html)

e. Pathway Section

Information	**Origin**
Pathways	GeneCards (https://www.genecards.org/)

f. Phenotype Section

Information	**Origin**
Disease ID	Downloaded database from DisGeNET (www.disgenet.org/)
Disease Name	Downloaded database from DisGeNET (www.disgenet.org/)
Source	Downloaded database from DisGeNET (www.disgenet.org/)

FIGURES

FIGURE 1. Workflow of RarePedia

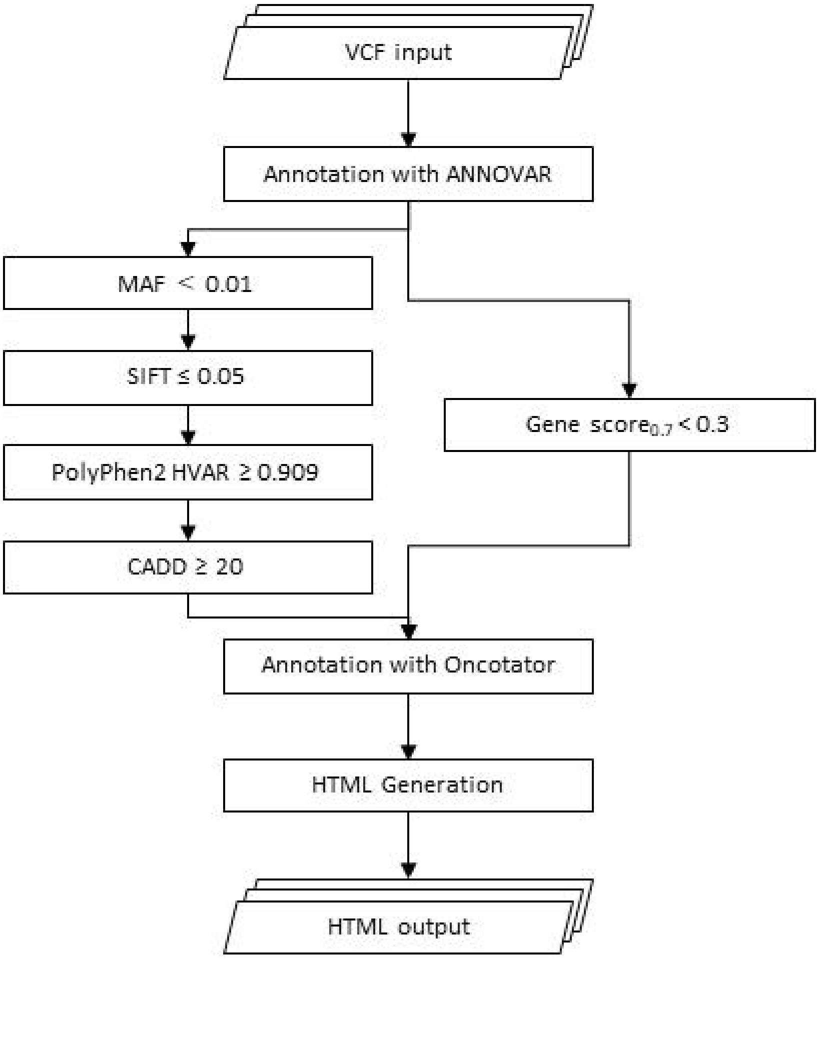


FIGURE 2. Statistics of rare-damaging variants from 1000 Genomes Project data

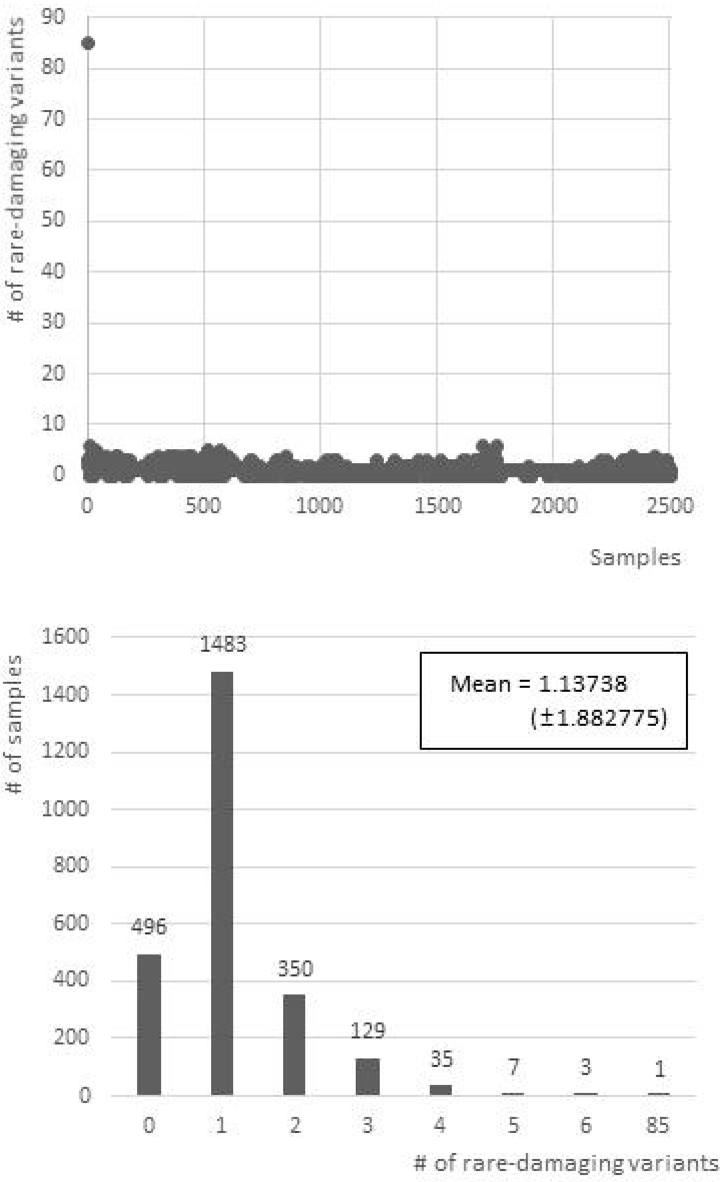
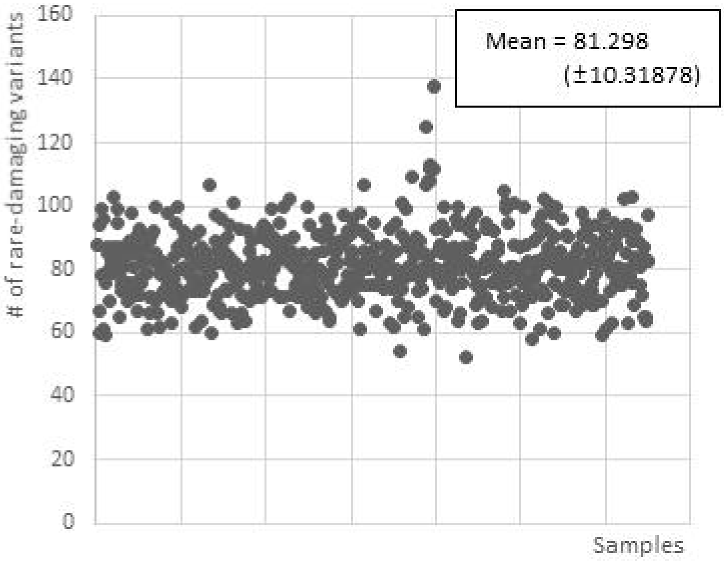


FIGURE 3. Statistics of rare-damaging variants from Alzheimer's Disease Sequencing Project (ADSP) data

a. Number of rare-damaging variants in each EOAD samples



b. Number of rare-damaging variants with allele frequencies in each EOAD samples

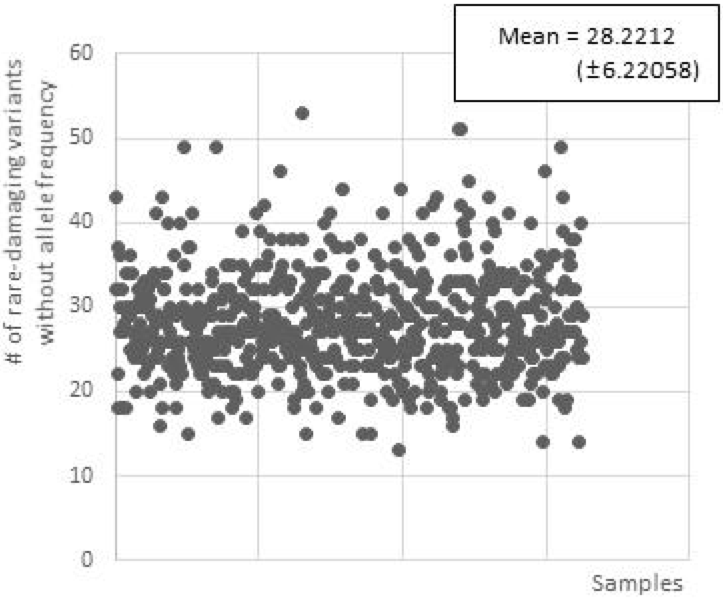
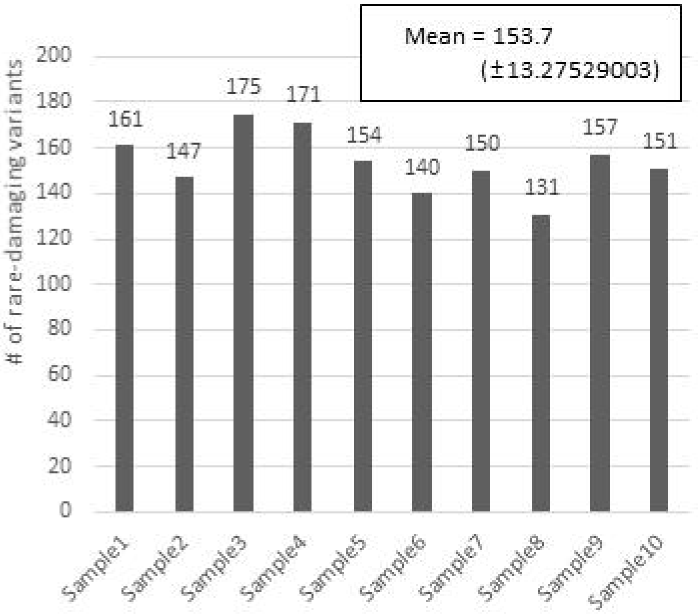


FIGURE 4. Statistics of rare-damaging variants from Korean Whole-Exome Sequencing data

a. Number of rare-damaging variants in each samples



b. Number of rare-damaging variants with allele frequencies in each samples

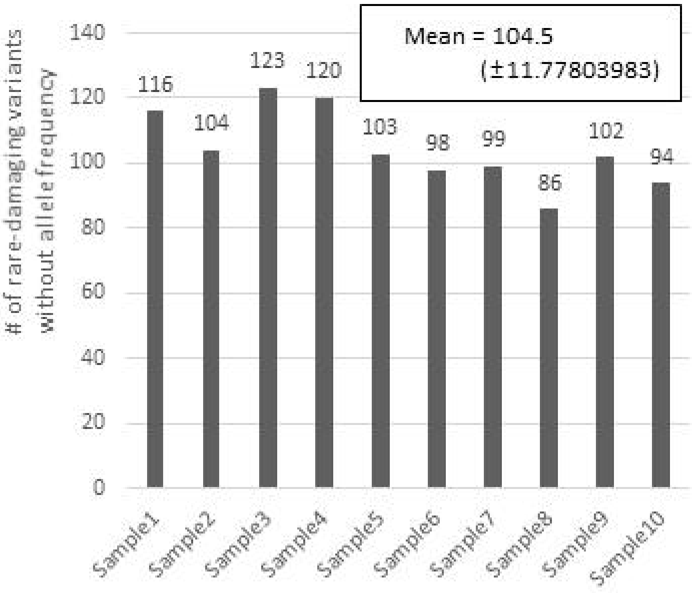
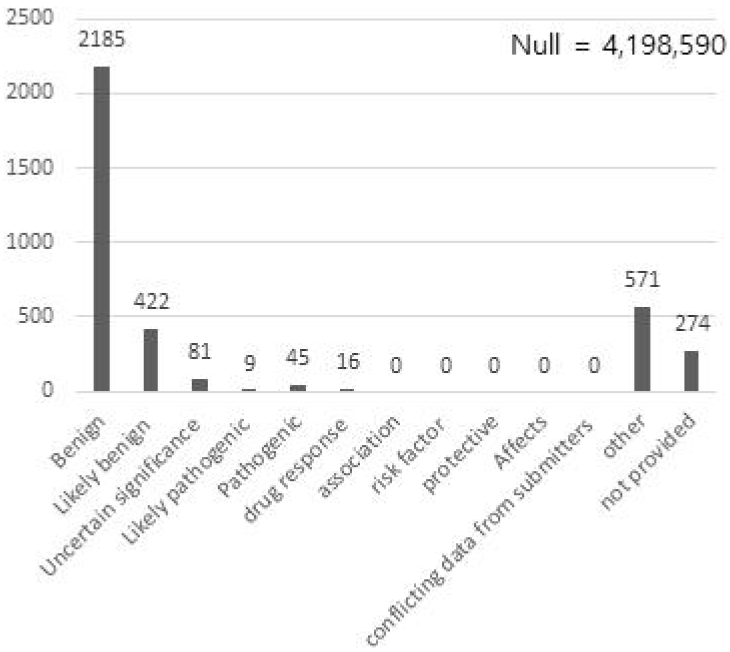
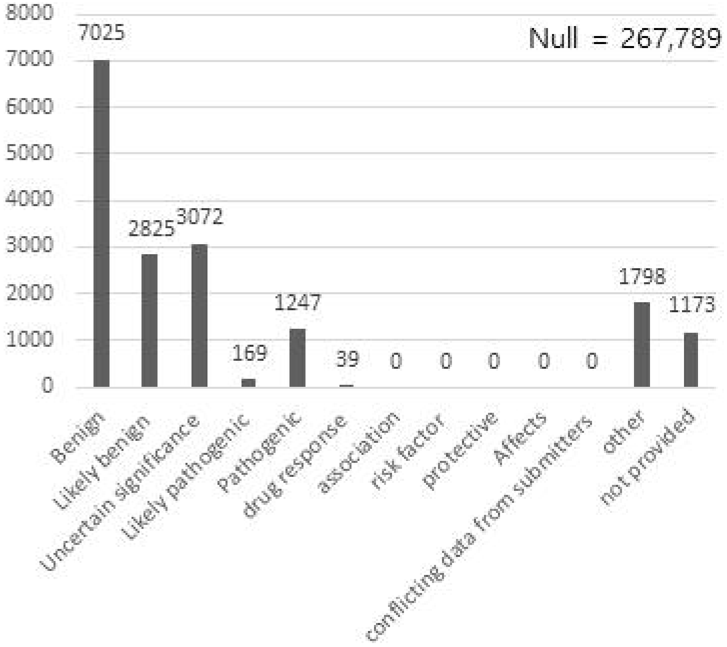


FIGURE 5. Statistics of ClinVar variants in three datasets

a. ClinVar variants in 1000 Genomes Project data



b. ClinVar variants in Alzheimer's Disease Sequencing Project(ADSP) data



c. ClinVar variants in 10 Korean whole exome sequencing data

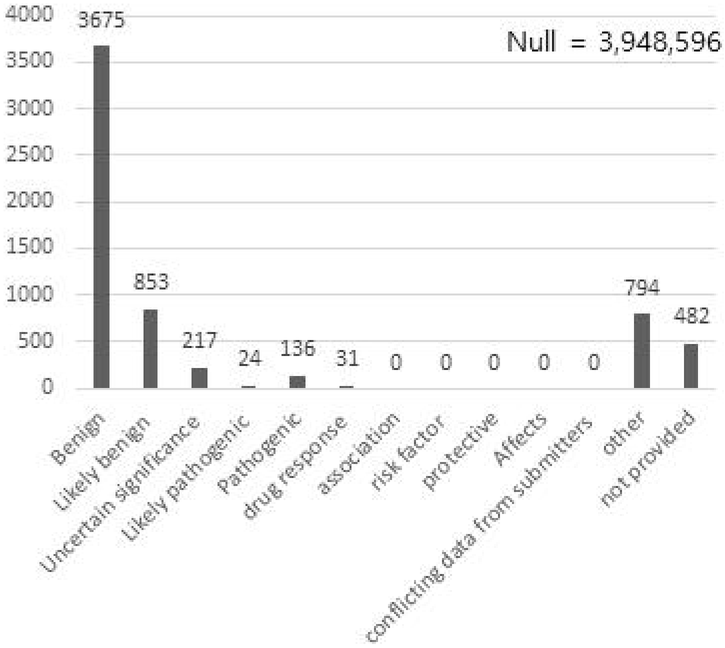
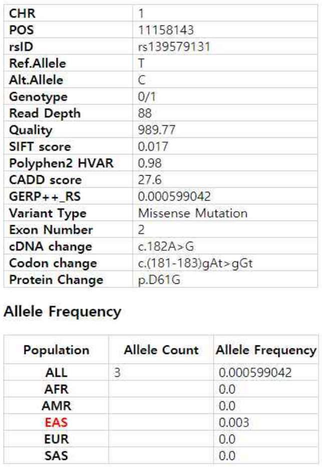


FIGURE 6. An example HTML page of one rare-damaging variant

a. Variant Section

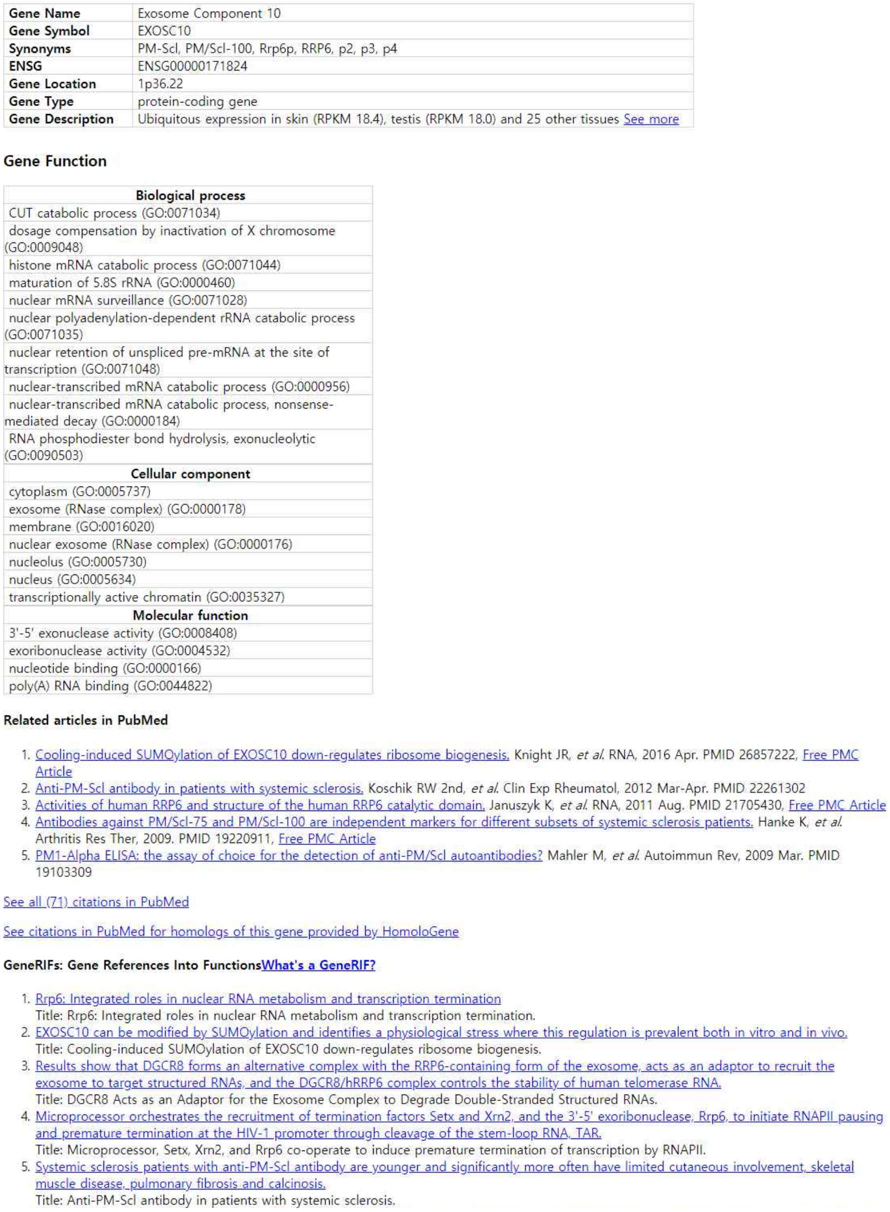


CHR	1
POS	11158143
rsID	rs139579131
Ref.Allele	T
Alt.Allele	C
Genotype	0/1
Read Depth	88
Quality	989.77
SIFT score	0.017
Polyphen2 HVAR	0.98
CADD score	27.6
GERP++_RS	0.000599042
Variant Type	Missense Mutation
Exon Number	2
cDNA change	c.182A>G
Codon change	c.(181-183)gAt>gGt
Protein Change	p.D61G

Allele Frequency

Population	Allele Count	Allele Frequency
ALL	3	0.000599042
AFR		0.0
AMR		0.0
EAS		0.003
EUR		0.0
SAS		0.0

b. Gene Section



Gene Name	Exosome Component 10
Gene Symbol	EXOSC10
Synonyms	PM-Scl, PM/Scl-100, Rrp6p, RRP6, p2, p3, p4
ENSG	ENSG00000171824
Gene Location	1p36.22
Gene Type	protein-coding gene
Gene Description	Ubiquitous expression in skin (RPKM 18.4), testis (RPKM 18.0) and 25 other tissues See more

Gene Function

Biological process
CUT catabolic process (GO:0071034)
dosage compensation by inactivation of X chromosome (GO:0009048)
histone mRNA catabolic process (GO:0071044)
maturation of 5.8S rRNA (GO:0000460)
nuclear mRNA surveillance (GO:0071028)
nuclear polyadenylation-dependent rRNA catabolic process (GO:0071035)
nuclear retention of unspliced pre-mRNA at the site of transcription (GO:0071048)
nuclear-transcribed mRNA catabolic process (GO:0000956)
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay (GO:0000184)
RNA phosphodiester bond hydrolysis, exonucleolytic (GO:0090503)
Cellular component
cytoplasm (GO:0005737)
exosome (RNase complex) (GO:0000178)
membrane (GO:0016020)
nuclear exosome (RNase complex) (GO:0000176)
nucleolus (GO:0005730)
nucleus (GO:0005634)
transcriptionally active chromatin (GO:0035327)
Molecular function
3'-5' exonuclease activity (GO:0008408)
exoribonuclease activity (GO:0004532)
nucleotide binding (GO:0000166)
poly(A) RNA binding (GO:0044822)

Related articles in PubMed

1. Cooling-induced SUMOylation of EXOSC10 down-regulates ribosome biogenesis. Knight JR, *et al.* RNA, 2016 Apr. PMID 26857222, Free PMC Article
2. Anti-PM-Scl antibody in patients with systemic sclerosis. Koschik RW 2nd, *et al.* Clin Exp Rheumatol, 2012 Mar-Apr. PMID 22261302
3. Activities of human RRP6 and structure of the human RRP6 catalytic domain. Januszyk K, *et al.* RNA, 2011 Aug. PMID 21705430, Free PMC Article
4. Antibodies against PM/Scl-75 and PM/Scl-100 are independent markers for different subsets of systemic sclerosis patients. Hanke K, *et al.* Arthritis Res Ther, 2009. PMID 19220911, Free PMC Article
5. PM1-Alpha ELISA: the assay of choice for the detection of anti-PM/Scl autoantibodies? Mahler M, *et al.* Autoimmun Rev, 2009 Mar. PMID 19103309

See all (71) citations in PubMed

See citations in PubMed for homologs of this gene provided by HomoloGene

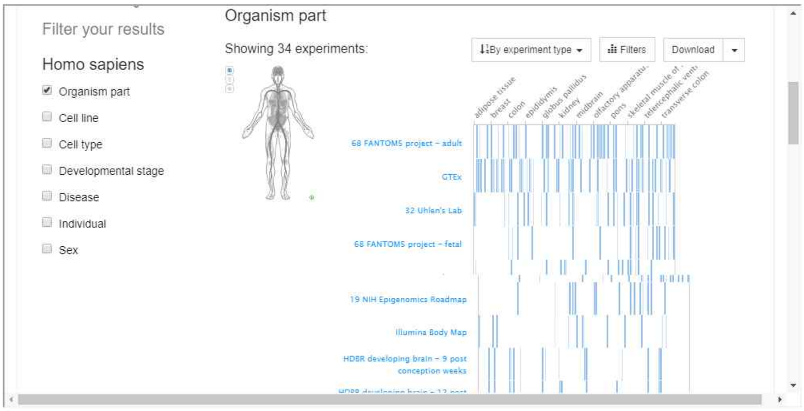
GeneRIFs: Gene References Into FunctionsWhat's a GeneRIF?

1. Rrp6: Integrated roles in nuclear RNA metabolism and transcription termination
   Title: Rrp6: Integrated roles in nuclear RNA metabolism and transcription termination.
2. EXOSC10 can be modified by SUMOylation and identifies a physiological stress where this regulation is prevalent both in vitro and in vivo.
   Title: Cooling-induced SUMOylation of EXOSC10 down-regulates ribosome biogenesis.
3. Results show that DGCR8 forms an alternative complex with the RRP6-containing form of the exosome, acts as an adaptor to recruit the exosome to target structured RNAs, and the DGCR8/hRRP6 complex controls the stability of human telomerase RNA.
   Title: DGCR8 Acts as an Adaptor for the Exosome Complex to Degrade Double-Stranded Structured RNAs.
4. Microprocessor orchestrates the recruitment of termination factors Setx and Xrn2, and the 3'-5' exoribonuclease, Rrp6, to initiate RNAPII pausing and premature termination at the HIV-1 promoter through cleavage of the stem-loop RNA, TAR.
   Title: Microprocessor, Setx, Xrn2, and Rrp6 co-operate to induce premature termination of transcription by RNAPII.
5. Systemic sclerosis patients with anti-PM-Scl antibody are younger and significantly more often have limited cutaneous involvement, skeletal muscle disease, pulmonary fibrosis and calcinosis.
   Title: Anti-PM-Scl antibody in patients with systemic sclerosis.

c. Transcript Section

Ensembl transcript ID	ENST00000376936.4
Transcript position	231
Transcript Strand	-
Refseq mRNA Id	NM_001001998.1

Baseline expression



Filter your results

Homo sapiens

☑ Organism part

□ Cell line

□ Cell type

□ Developmental stage

□ Disease

□ Individual

□ Sex

Organism part

Showing 34 experiments:

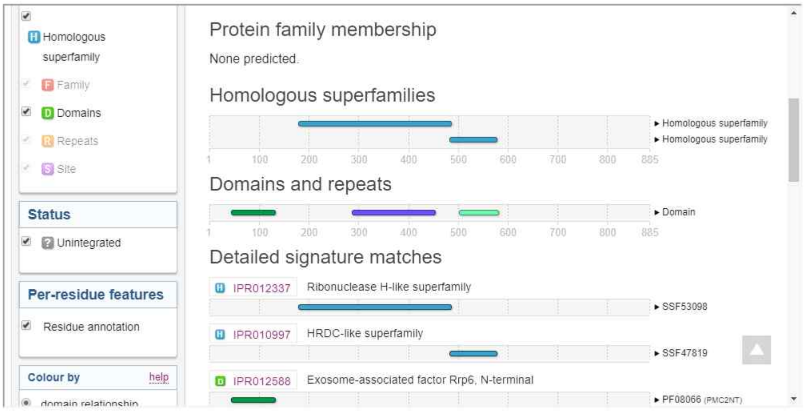
Expression in other tissues OtherTissue

Expression in cancer tissues CancerTissue

d. Protein Section

Recommended name	Exosome Component 10
UniProtKB/Swiss-Prot	Q01780
Ensembl proteins	ENSP00000366135
RefSeq protein	NP_001001998.1
Protein size	**Mass :** 885 **Da :** 100,831

Protein family



Protein family membership

None predicted.

Homologous superfamilies

Domains and repeats

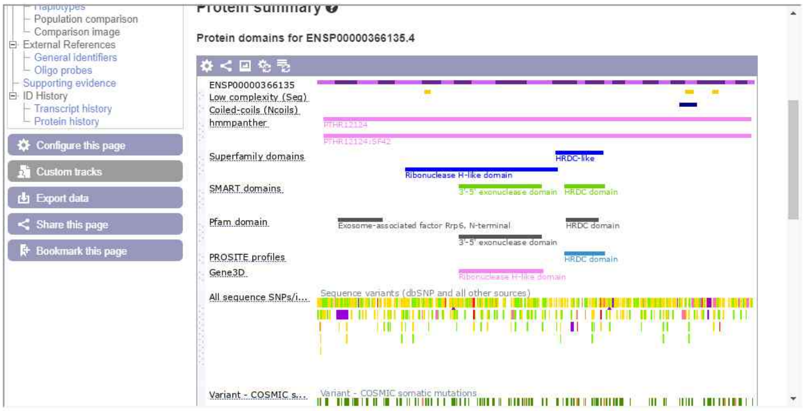
Detailed signature matches

IPR012337 Ribonuclease H-like superfamily ▸ SSF53098

IPR010997 HRDC-like superfamily ▸ SSF47819

IPR012588 Exosome-associated factor Rrp6, N-terminal ▸ PF08066 (PMC2NT)

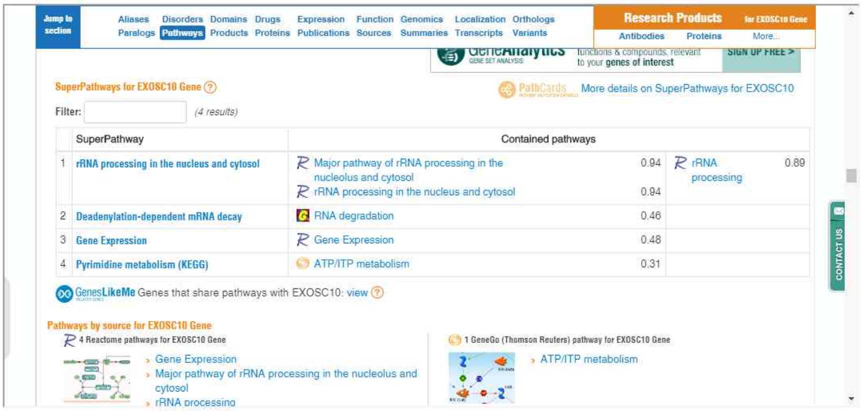
Protein damain



Protein summary

Protein domains for ENSP00000366135.4

e. Pathway Section



SuperPathways for EXOSC10 Gene ?

More details on SuperPathways for EXOSC10

Filter: (4 results)

	SuperPathway	Contained pathways			
1	rRNA processing in the nucleus and cytosol	Major pathway of rRNA processing in the nucleolus and cytosol rRNA processing in the nucleus and cytosol	0.94 0.94	rRNA processing	0.89
2	Deadenylation-dependent mRNA decay	RNA degradation	0.46		
3	Gene Expression	Gene Expression	0.48		
4	Pyrimidine metabolism (KEGG)	ATP/ITP metabolism	0.31		

GenesLikeMe Genes that share pathways with EXOSC10: view ?

Pathways by source for EXOSC10 Gene

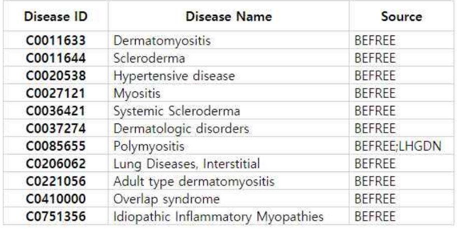
4 Reactome pathways for EXOSC10 Gene

- Gene Expression
- Major pathway of rRNA processing in the nucleolus and cytosol
- rRNA processing

1 GeneGo (Thomson Reuters) pathway for EXOSC10 Gene

- ATP/ITP metabolism

f. Phenotyep Section



Disease ID	Disease Name	Source
C0011633	Dermatomyositis	BEFREE
C0011644	Scleroderma	BEFREE
C0020538	Hypertensive disease	BEFREE
C0027121	Myositis	BEFREE
C0036421	Systemic Scleroderma	BEFREE
C0037274	Dermatologic disorders	BEFREE
C0085655	Polymyositis	BEFREE;LHGDN
C0206062	Lung Diseases, Interstitial	BEFREE
C0221056	Adult type dermatomyositis	BEFREE
C0410000	Overlap syndrome	BEFREE
C0751356	Idiopathic Inflammatory Myopathies	BEFREE

국문초록

질병의 원인이 되는 인자를 찾기 위한 연구원들의 노력은 오래 전부터 계속되고 있다. 환경적인 요인, 스트레스, 노화가 주된 원인으로 손꼽혀 왔다. 과학 기술의 발전으로, 많은 질병 연관 유전자와 기작이 밝혀졌다. DNA 염기 서열 분석 기술의 발전과 함께, 질병과 관련된 유전자들의 염기 서열, 심지어 한 유전자 내에 존재하는 하나의 특정 질병 연관 유전적 변이까지도 밝혀지고 있다. 하지만, 매우 낮은 빈도로 발생하는 변이들의 경우, 분석에서 무시되는 경우가 많다. 이러한 희귀 변이들에 대한 정보가 적다는 것, 그리고 표본 크기를 늘리지 않고서는 이들 변이에 대한 정확한 분석이 어렵다는 것이 그 이유이다. 다시 말해, 특정 표현형과 연관이 있다고 단정 짓기에는 많은 어려움이 있어, 배제할 수 밖에 없는 경우가 많다는 것이다.

레어피디아는 희귀 변이에 대한 정보를 통합하기 위해 고안되었다. 더 나아가, 질병과 연관이 있을 것으로 예측되는 손상 변이에 초점을 맞추고 있다. 레어피디아는 차세대 염기 서열 분석 데이터를 이용하는 또 하나의 새로운 방법이며, 여러 곳에 산재하여 있는 정보를 동시에 볼 수 있도록 도와주는 도구이다.

레어피디아의 궁극적인 목표는 이러한 희귀-손상 변이에 대한 새로운 정보가 있는 경우, 추가적인 업데이트를 통한 정보의 축적이다. 지금까지 체계적으로 정리되지 못하였던 희귀-손상 변이에 대한 정보를 정리할

수 있는 수단이 될 수 있을 것이다.

주요어 : 희귀-손상 변이, 유전자, 차세대 염기서열 분석, 방법론

학 번 : 2016-20463