Genetics inMedicine | BRIEF REPORT

Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification

Wei Song, PhD[1,2], Sabrina A. Gardner, PhD[1,2], Hayk Hovhannisyan, PhD[1,2], Amanda Natalizio, PhD[1,2], Katelyn S. Weymouth, PhD[1,2], Wenjie Chen, PhD[1,2], Ildiko Thibodeau, PhD[1,2], Ekaterina Bogdanova, PhD[1,2], Stanley Letovsky, PhD[1,2], Alecia Willis, PhD, FACMG[1,2] and Narasimhan Nagan, PhD, FACMG[1,2]

Purpose: We evaluated the Exome Aggregation Consortium (ExAC) database as a control cohort to classify variants across a diverse set of genes spanning dominant and recessively inherited disorders.

Methods: The frequency of pathogenic variants in ExAC was compared with the estimated maximal pathogenic allele frequency (MPAF), based on the disease prevalence, penetrance, inheritance, allelic and locus heterogeneity of each gene. Additionally, the observed carrier frequency and the ethnicity-specific variant distribution were compared between ExAC and the published literature.

Results: The carrier frequency and ethnic distribution of pathogenic variants in ExAC were concordant with reported estimates. Of 871 pathogenic/likely pathogenic variants across 19 genes, only 3 exceeded the estimated MPAF. Eighty-four percent of variants with ExAC frequencies above the estimated MPAF were classified as "benign." Additionally, 20% of the cardiac and 19% of the Lynch syndrome gene variants originally classified as "VUS" occurred with ExAC frequencies above the estimated MPAF, making these suitable for reassessment.

Conclusions: The ExAC database is a useful source for variant classification and is not overrepresented for pathogenic variants in the genes evaluated. However, the mutational spectrum, pseudogenes, genetic heterogeneity, and paucity of literature should be considered in deriving meaningful classifications using ExAC.

Genet Med advance online publication 17 December 2015

Key Words: clinical testing; databases; genetic testing; population genetics; variant classification

INTRODUCTION

Rapid advances in sequencing technologies have resulted in increasingly more genetic testing services, ranging from single-gene analysis to targeted panels and whole-exome and whole-genome sequencing. In clinical settings, the limiting factor has shifted from acquisition of sequencing data to classification, interpretation, and reporting of novel and recurring sequence variants with little or no conclusive information supporting causation.[1]

Classification of sequence variants considers the prevalence of the variant in presumably healthy unaffected individuals, cosegregation of the variant with disease in families, and computational and in vitro/in vivo analyses showing the predicted effect of the variant on function or aberrant splicing.[2] In particular, the frequency of occurrence, or lack thereof, of a variant in the general population (controls) constitutes an important line of evidence impacting variant classification. Additionally, these databases are utilized in next-generation sequencing pipelines to exclude common variants that are less likely to be pathogenic.[3,4] If the frequency threshold is set too low or if the data set used to ascertain frequency contains affected individuals, then potentially disease-causing variants may be filtered out in the early stages of the pipeline. Therefore, the utility of large frequency databases to support classification and analysis of variants is rapidly gaining momentum.

The Exome Aggregation Consortium (ExAC),[5] a collection of whole-exome sequencing data from more than 60,000 ostensibly healthy individuals representing diverse human populations, was released in late 2014. The aim of this study was to evaluate this database as a representative control cohort for analysis and classification of sequence variants observed in a clinical laboratory. In particular, we wanted to explore whether the ExAC data set was enriched for pathogenic variation in specific disorders or genes. As the number, diversity, and heterogeneity of genes and disorders tested in clinical settings are rather diverse, we decided to pilot our study to include a broad, but representative, sampling of dominant tumor suppressor genes, dominant cardiovascular-disorder genes, and recessive genes with well-established clinical utility and uptake in clinical diagnostic settings.

The first two authors contributed equally to this work.

[1]Integrated Genetics, Laboratory Corporation of America® Holdings, Westborough, Massachusetts, USA; [2]Integrated Genetics, Laboratory Corporation of America® Holdings, Research Triangle Park, North Carolina, USA. Correspondence: Narasimhan Nagan (Narasimhan.Nagan@integratedgenetics.com)

Submitted 9 September 2015; accepted 30 October 2015; advance online publication 17 December 2015. doi:10.1038/gim.2015.180

BRIEF REPORT

MATERIALS AND METHODS

Data collection and analysis

The ExAC data set provides sequence variation in 60,706 unrelated individuals from various disease-specific and population genetic studies. The data set includes a distribution of diverse ethnicities including European (non-Finnish), European (Finnish), African, Latino, South Asian, East Asian, and "Other." Sequencing data from 17 contributing projects were included in ExAC. Although phenotype data for the individuals included have not been provided, individuals affected by severe pediatric disease were excluded from the data set.

Variants for analysis comprised a collection of our internal classifications in 19 genes to include dominant tumor suppressor genes (*BRCA1*, *BRCA2*, *MLH1*, *MSH2*, *MSH6*, and *PMS2*), dominant cardiac-disorder genes (*MYBPC3*, *MYH7*, *TNNT2*, *TNNI3*, *PKP2*, *DSG2*, *DSP*, *DSC2*, and *FBN1*), and recessive genes (*CFTR*, *GJB2*, *HBB*, and *MEFV*). All variants were classified by a tiered in-house variant classification protocol (https://submit.ncbi.nlm.nih.gov/ft/byid/pttb9itm/labcorp_variant_classification_method_-_may_2015.pdf) following guidelines issued by the American College of Medical Genetics and Genomics (ACMG).[2] The data presented encompass 2,984 classified variants across 19 genes spanning diverse disorders and modes of inheritance.

ExAC data for each gene were downloaded from http://exac.broadinstitute.org/.[5] The corresponding frequency in ExAC of each variant in the data set described above was queried by the corresponding nucleotide level nomenclature scheme (c.name). Differences in nomenclature between ExAC and our internal variant database were reconciled with the HGVS-approved standard for each variant in the data set to ensure accuracy of ascertainment.[6]

To derive traceable comparisons for each gene, the evidence supporting phenotype prevalence, locus/allelic heterogeneity, and penetrance was used to estimate the maximal pathogenic allele frequency (MPAF) for each gene (**Supplementary Table S1** online). MPAF provides a conservative maximum expected frequency of pathogenic alleles in any gene under the assumption that the corresponding disease is entirely attributable to a single pathogenic variant.[7] Variants present at frequencies above MPAF provide supportive evidence for nonpathogenicity.

For each gene, the frequency in ExAC was determined for all variants classified as pathogenic or likely pathogenic. In addition, all classified variants with frequencies above the MPAF in each gene were ascertained. Carrier frequencies and ethnicity-specific variant distribution(s) in ExAC were compared with the published literature for variants in genes with available information.

RESULTS

The three pathogenic *BRCA* variants with the highest allele frequency in ExAC were the well-known Ashkenazi Jewish (AJ) founder mutations, namely, *BRCA2* c.5946delT (32/120,698 chromosomes), *BRCA1* c.68_69delAG (29/120,972 chromosomes), and *BRCA1* c.5266dupC (19/121,412 chromosomes). The carrier frequency of 1/756 for the three *BRCA1* and *2* AJ founder mutations in the ExAC database was consistent with the frequency of 1 in 400 to 800 individuals reported to carry pathogenic germ-line mutations in *BRCA1 or BRCA2* in the general population.[8–10]

The carrier frequency of the most frequent AJ mutation, c.3846G>A (p.W1282X) in ExAC was 1/1312, which is lower than the reported carrier frequency of 1/863 for this *CFTR* variant in an ethnically diverse US population ($P < 0.05$).[11] This indicates that the AJ ethnicity is not overrepresented in the ExAC data set. Likewise, the three most frequent pathogenic *CFTR* variants observed in ExAC were c.1521_1523delCTT (p.F508del), c.350G>A (p.R117H), and c.3209G>A (p.R1070Q), each with a carrier frequency of 1/74, 1/325, and 1/619, respectively. Of these, the carrier frequency of p.F508del and p.R117H in ExAC were in range of the reported frequency for p.F508del (1/65) and p.R117H (1/422) in an ethnically diverse US population.[11] Within the subpopulations represented in ExAC, the carrier frequencies of these three most frequent pathogenic *CFTR* variants are highest in non-Finnish Europeans (1/47) for p.F508del, in non-Finnish Europeans (1/195) for p.R117H, and in South Asians (1/95) for p.R1070Q. The overall distribution pattern of these variants within different ethnicities is consistent with published data among African, Asian, Caucasian, Latino, and other populations.[11,12] Furthermore, the distribution of pathogenic variants with homozygous occurrences in *GJB2* (p.V37I and c.35delG in East Asians and Europeans, respectively), *HBB* (p.E7K and p.E7V in Africans), and *MEFV* (p.V726A in Europeans) followed the expected distribution based upon the reported prevalence of autosomal recessive deafness (*GJB2*, OMIM 220290), hemoglobinopathies (*HBB*, OMIM 141900), and Familial Mediterranean Fever (*MEFV*, OMIM 249100) in these subpopulations.[13–15] These observations demonstrate that ExAC is not overenriched for pathogenic variants in the specific disorders tested, thereby supporting its utility as a control cohort in genetic analysis.

Only 3 of 871 variants (0.34%) that had been classified as pathogenic or likely pathogenic across 19 genes exceeded the estimated MPAF. The distribution in ExAC of the average minor allele frequency (MAF) of pathogenic and likely pathogenic variants in relation to the corresponding estimated MPAF in the genes analyzed is provided in **Table 1**.

Of 237 *BRCA1* and *2* variants that have been classified as pathogenic or likely pathogenic, 44 were present in ExAC. The majority of these variants had an allele count of 1 or 2 of about 121,412 total chromosomes ($n = 35$). None had an allele frequency exceeding the MPAF for each gene.

Of the 266 cardiac-disorder gene variants that have been classified as pathogenic or likely pathogenic, 32 were present in ExAC. The majority of these variants had an allele count of 1 or 2 of about 121,412 total chromosomes ($n = 20$). Three variants, *DSG2,* c.1174G>A (p.Val392Ile), *TNNT2*, c.832C>T (p.Arg278Cys) and *PKP2*, c.419C>T (p.Ser140Phe), had a frequency that exceeded the MPAF for a pathogenic variant by 10-, 3-, and 4-fold, respectively. Each of these has been

reevaluated by our laboratory with the *DSG2* and *PKP2* variants being reclassified as likely benign and the *TNNT2* variant being reclassified as VUS.

Of 87 Lynch syndrome (OMIM 120435) variants that have been classified as pathogenic or likely pathogenic, 14 were present in ExAC. None of these variants had an allele frequency exceeding the MPAF for each gene.

For genes associated with recessively inherited disorders, namely *CFTR*, *GJB2*, *HBB*, and *MEFV*, a total of 133 variants that have been classified as pathogenic or likely pathogenic were present in ExAC. As with breast cancer and Lynch syndrome genes, none of these variants had an allele frequency exceeding the MPAF for each gene.

Eighty-four percent of variants with frequencies above the MPAF in ExAC were classified as "benign/likely benign" (Table 2). Additionally, 20% of cardiac and 19% of Lynch syndrome gene variants originally classified as "VUS"(variant of uncertain clinical significance) occurred with ExAC frequencies above the estimated MPAF, making these worthy of reassessment.

DISCUSSION

The use of the estimated MPAF for each gene illustrated in this study represents a traceable paradigm for assessing the impact of variant occurrences in population databases as supportive evidence of non-pathogenicity. As demonstrated with *BRCA* and *CFTR*, the carrier frequency and ethnicity-specific distribution of classic, well-studied pathogenic variants in our data set matched the values reported from the general population and it was not overrepresented by variation specific to ethnicities

Table 1 Average MAF of pathogenic variants in analyzed genes

Gene	Number of variants	Number of pathogenic/likely pathogenic	Number of pathogenic with MAF >0	Number of pathogenic above MPAF	MPAF	Average MAF for pathogenic variants[a]	Range MAF for pathogenic variants[a]
BRCA1	385	110	24	0	0.001	0.000031381	0.000008238 to 0.0002397
BRCA2	615	127	20	0	0.00075	0.000025368	0.000008242 to 0.0002651
MYPBC3	154	59	10	0	0.001	0.000039601	0.000008293 to 0.0002027
MYH7	124	30	8	0	0.00125	0.000014424	0.000008237 to 0.00003296
TNNT2	27	8	3	1[b]	0.0005	0.000151469	0.000008238 to 0.0004291
TNNI3	23	3	2	0	0.000125	0.00001664	0.000008299 to 0.00002498
PKP2	35	10	4	1[b]	0.0005375	0.000600598	0.00001647 to 0.00232
DSG2	31	2	2	1[b]	0.000125	0.001287643	0.000008286 to 0.002567
DSP	59	5	0	0	0.0002	-	-
DSC2	25	1	0	0	0.0000625	-	-
FBN1	489	148	3	0	0.0001125	0.000016488	0.000008238 to 0.00003298
MLH1	66	21	0	0	0.00071	-	-
MSH2	68	29	2	0	0.000568	0.0000082405	0.000008237 to 0.000008244
MSH6	77	24	3	0	0.000142	0.00000825433	0.000008242 to 0.000008268
PMS2	60	13	9	0	0.000114	0.0000132703	0.00000824 to 0.00003295
CFTR	383	143	71	0	0.013	0.000208846	0.000008237 to 0.006785
GJB2	85	38	24	0	0.026	0.00068355	0.000008238 to 0.006587
HBB	203	87	28	0	0.0112	0.000324687	0.000008238 to 0.004384
MEFV	75	13	10	0	0.0217	0.000971982	0.00001647 to 0.005502
Total	2,984	871	223	3			

MAF, minor allele frequency; MPAF, maximal pathogenic allele frequency.

[a]Average MAF for those variants present in Exome Aggregation Consortium (having a MAF above 0). [b]Since the time of analysis these variants have been reclassified and are no longer classified as pathogenic.

Table 2 Classification of variants with MAF higher than MPAF in the analyzed gene sets

Gene	Number of variants above estimated MPAF	% above MPAF with classification of benign/likely benign	% above MPAF with classification of VUS
BRCA1&2	86	98%	2%
Cardiac genes	214	79%	20%
Lynch syndrome genes	75	81%	19%
Recessive genes	19	95%	5%
Total	407	84%	15%

MAF, minor allele frequency; MPAF, maximal pathogenic allele frequency.

such as the AJ. Therefore, ExAC is not enriched for pathogenic variation in the specific disorders and genes evaluated, making it a useful data set to facilitate accurate classification outcomes.

Next, we used ExAC occurrences to identify variants in our database that could be reclassified in light of new evidence. Only 3 of 871 variants originally classified as pathogenic or likely pathogenic were present in ExAC at frequencies exceeding the estimated MPAF. Each of these three variants was in a gene associated with inherited cardiac disorders and had been originally classified conservatively prior to the large population control databases such as ESP and ExAC. Therefore, ExAC served as useful supporting evidence to merit a reevaluation of the pathogenicity of these variants.

Lastly, a majority (84%) of variants that had frequencies above the estimated MPAF were appropriately classified as benign or likely benign. Specifically, 98% of variants in *BRCA1 and BRCA*2 genes and 95% of variants observed among the 4 genes associated with recessive disorders (*CFTR*, *GJB2*, *HBB*, and *MEFV*) that had frequencies above the estimated MPAF were classified as benign or likely benign. Variants in cardiac and Lynch syndrome genes were the two exceptions to this observation. Forty-three cardiac gene variants and 12 Lynch syndrome gene variants that were originally classified as VUS had an ExAC frequency exceeding the estimated MPAF. Ten of the 43 cardiac variants were found in ethnic groups that were not represented in ESP (Latino, East Asian, and South Asian), and they would not have been observed prior to the release of ExAC. Eighteen of the 43 cardiac gene variants had a frequency only one- to threefold above the estimated MPAF and could not be considered strong evidence for classification as benign. The remaining cardiac gene variants represent a subset associated with factors such as digenic inheritance, low penetrance, population specific variation, or potential role as disease modifiers, causing their classification to be conservative, even with significant occurrences of the variant in the control population.[16,17]

Nine of the 12 (75%) Lynch syndrome variants with an ExAC frequency exceeding the MPAF were in the *PMS2* gene. Analysis of variants in *PMS2* is challenging owing to the presence of numerous pseudogenes with high homology that preclude unequivocal differentiation between true variants versus those originating in the pseudogenes.[18–20] Because of a high rate of mismapping of next-generation sequencing alignments in pseudogene regions, reports that do not include long-range PCR or RNA analysis to specifically distinguish variant occurrence between the gene and pseudogene are not weighted in our classifications. ESP, 1000 genomes, and ExAC do not specifically rule out pseudogene interference, which makes them less useful. Therefore, *PMS2* variants present at high frequency with little supporting data are more likely be classified conservatively as a VUS. As with cardiac genes, the remaining 3 Lynch syndrome variants had a frequency of one- to twofold above the MPAF, not reaching a threshold for unequivocal classification as benign.

A limitation of ExAC is the use of non-HGVS standard variant nomenclature. This increases the likelihood of false negative observations. Although, single-nucleotide variants are likely to be called accurately, heightened awareness in reviewing the annotation of variants, such as deletions and insertions, is recommended. In conclusion, our observations support ExAC as a control cohort for classifying variants in clinical settings. We recommend that this database be evaluated across diverse sets of genes and disorders, mindful of underlying genetic complexities (such as pseudogenes) that pose challenges in deriving meaningful classifications using control data sets.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at http://www.nature.com/gim

ACKNOWLEDGMENTS

The authors thank the Exome Aggregation Consortium and the groups that provided exome variant data for comparison. A full list of contributing groups can be found at http://exac.broadinstitute.org/about.

DISCLOSURE

At the time this study was conducted, all authors were employed by Integrated Genetics, Laboratory Corporation of America® Holdings, and may hold stock of and/or stock options with LabCorp.

REFERENCES

1. Amendola LM, Dorschner MO, Robertson PD, et al. Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome Res* 2015;25:305–315.
2. Richards S, Aziz N, Bale S, et al.; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17:405–424.
3. Bao R, Huang L, Andrade J, et al. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform* 2014;13(suppl 2):67–82.
4. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* 2012;20:490–497.
5. Exome Aggregation Consortium (ExAC), Cambridge, MA. http://exac.broadinstitute.org. Accessed March 2015.
6. den Dunnen JT, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat* 2000;15:7–12.
7. Duzkale H, Shen J, McLaughlin H, et al. A systematic approach to assessing the clinical significance of genetic variants. *Clin Genet* 2013;84:453–463.
8. Ford D, Easton DF, Bishop DT, Narod SA, Goldgar DE. Risks of cancer in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. *Lancet* 1994;343:692–695.
9. Claus EB, Schildkraut JM, Thompson WD, Risch NJ. The genetic attributable risk of breast and ovarian cancer. *Cancer* 1996;77:2318–2324.
10. Whittemore AS, Gong G, Itnyre J. Prevalence and contribution of BRCA1 mutations in breast cancer and ovarian cancer: results from three U.S. population-based case-control studies of ovarian cancer. *Am J Hum Genet* 1997;60:496–504.
11. Rohlfs EM, Zhou Z, Heim RA, et al. Cystic fibrosis carrier testing in an ethnically diverse US population. *Clin Chem* 2011;57:841–848.
12. Abecasis GR, Auton A, Brooks LD, et al.; 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;491:56–65.
13. Ben-Chetrit E, Touitou I. Familial mediterranean Fever in the world. *Arthritis Rheum* 2009;61:1447–1453.
14. Chan DK, Chang KW. GJB2-associated hearing loss: systematic review of worldwide prevalence, genotype, and auditory phenotype. *Laryngoscope* 2014;124:E34–E53.

15. Kohne E. Hemoglobinopathies: clinical manifestations, diagnosis, and treatment. *Dtsch Arztebl Int* 2011;108:532–540.
16. Wessels MW, Herkert JC, Frohn-Mulder IM, et al. Compound heterozygous or homozygous truncating MYBPC3 mutations cause lethal cardiomyopathy with features of noncompaction and septal defects. *Eur J Hum Genet* 2015;23:922–928.
17. Westenskow P, Splawski I, Timothy KW, Keating MT, Sanguinetti MC. Compound mutations: a common cause of severe long-QT syndrome. *Circulation* 2004;109:1834–1841.
18. Hegde M, Ferber M, Mao R, Samowitz W, Ganguly A; Working Group of the American College of Medical Genetics and Genomics (ACMG) Laboratory Quality Assurance Committee. ACMG technical standards and guidelines for genetic testing for inherited colorectal cancer (Lynch syndrome, familial adenomatous polyposis, and MYH-associated polyposis). *Genet Med* 2014;16:101–116.
19. Thompson BA, Spurdle AB, Plazzer JP, et al.; InSiGHT. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nat Genet* 2014;46:107–115.
20. Lynch HT, Snyder CL, Shaw TG, Heinen CD, Hitchins MP. Milestones of Lynch syndrome: 1895-2015. *Nat Rev Cancer* 2015;15:181–194.