Variant reclassification based on the allele frequency information from exome sequencing data of 20,455 patients enriched with underrepresented populations

Kisang Kwon[1], Seong-In Hyun[1], Heonjong Han[1], Go Hun Seo[1], Hane Lee[1]

[1] 3billion, Inc, Seoul, South Korea

INTRODUCTION

- Accurate variant classification is essential for making a molecular diagnosis, and therefore, multiple lines of evidence are carefully considered for determining variant pathogenicity. One of the most important lines of evidence used for diagnosing rare Mendelian disorders is the minor allele frequency (AF) information in the population. Filtering out common variants based on the AF information is the most effective way of narrowing down the number of variants that need interpretation, as it removes a substantial number of variants that are too common to be disease-causing.
- There are several publicly available databases (DB) such as gnomAD that provide AF information from large populations. Many of these large DBs tend to be dominated by few populations though and therefore it is less effective when filtering out variants for individuals from under-represented populations.
- Here, we investigated how many of our internal variants from 20,455 exomes that are not part of any publicly available DBs can be classified as likely benign (LB), solely based on the AF information from our internal DB but not from gnomAD. A subset of these variants were present in ClinVar as pathogenic, likely pathogenic or VUS (P/LP/VUS) variants and we were able to downgrade them to likely benign (LB). An internally developed ethnicity predicting algorithm which estimates ethnicity based on the variant frequency information was used to confirm that the variants that can be reclassified by this method are indeed from underrepresented populations.

METHODS

All internal variants were queried to select those that can be classified as likely benign based on the allele frequency information. Variant selection steps are described below.

- **Step 1**: Extract all high-quality variants from internal variant DB for 20,455 samples
- **Step 2**: Select variants based on the disease inheritance pattern as follows:
  - Autosomal dominant (AD) genes: gnomAD allele count = 0
  - Autosomal recessive (AR) genes: gnomAD homozygous count = 0
  - X linked (XL) inheritance genes (including X-linked dominant genes): gnomAD allele count = 0
- **Step 3**: Remove variants that were reported as diagnoses. To avoid over-calling variants as LB for genes/disorders that are extremely rare or likely associated with incomplete penetrance, variable expressivity or late onset, only the genes that have >9 ClinVar SCVs and the variants with internal AF greater than the gnomAD allele frequency of the most common ClinVar P/LP variant for each genes were further selected.
- **Step 4**: Intersect with ClinVar P/LP/VUS variants.
- **Step 5**: Manually review the variants to determine if the variant can be reclassified as LB.

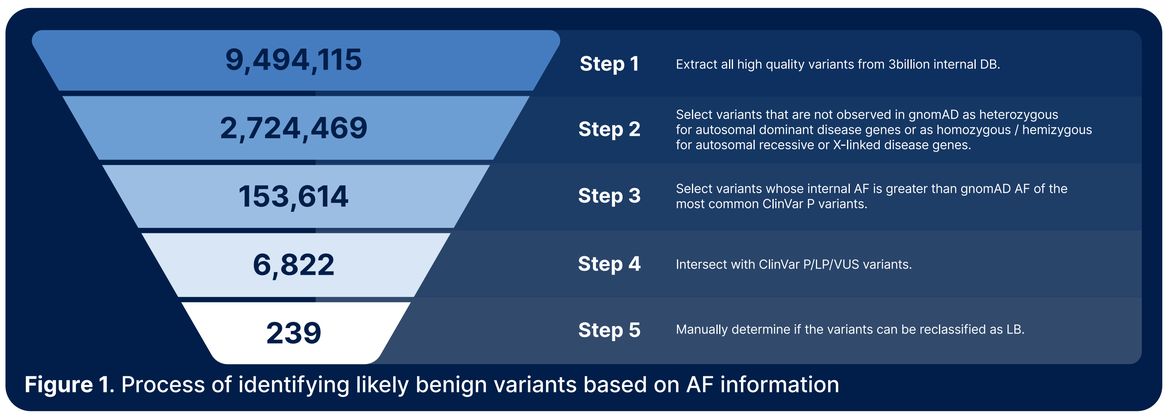

Figure 1. Process of identifying likely benign variants based on AF information

For each variant reclassified as LB, the ethnicity information was reviewed. Ethnicity for each sample was predicted based on the variant frequency information using ethnicity information in gnomAD as described below.

- The biological assumption of this method is that samples from the same ethnic group will have a similar set of variants.
- A variant frequency table in which the row is the common variant observed in gnomAD, and the column is the distribution of the ethnic origin of each variant represented in gnomAD was built. Common variants were variants meeting both of the following conditions: 1) 0.05 < allele frequency < 0.95 and 2) Total allele number > 2,000 and allele number in each ethnicity > 100
- To estimate the ethnicity score for each sample, the conditional probability a variant group would have occured from each ethnic group was calculated and normalized by geometric mean.

$$\textit{Ethnicity score} = \{\Pr(\text{Variants} \mid \textit{Ethnicity})\}^{\frac{1}{n}} = \left(\prod_{x=1}^{n} \Pr(\text{Variant}_x \mid \textit{Ethnicity})\right)^{\frac{1}{n}} = \left(\prod_{x=1}^{n} AF_x\right)^{\frac{1}{n}}$$

- Finally, for each sample, an ethnicity was assigned based on the highest ethnicity score calculated.

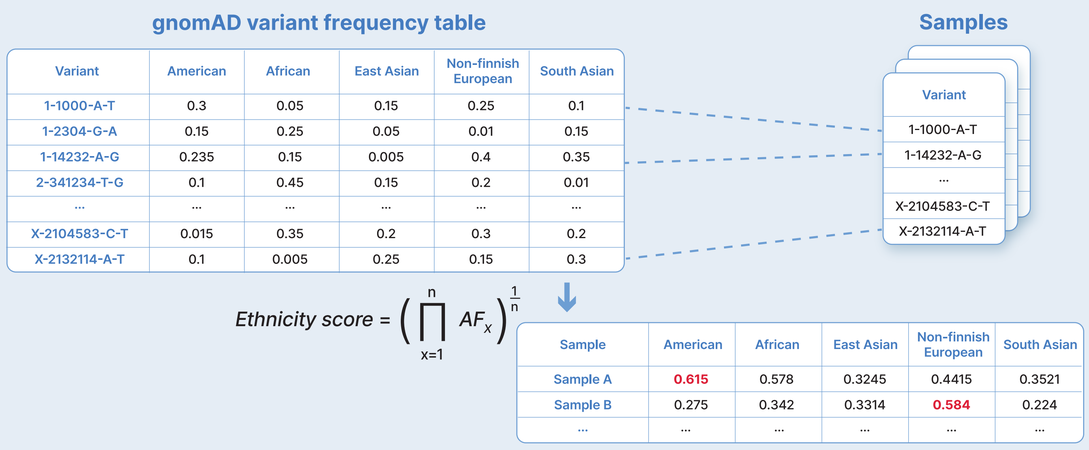

Figure 2. An overview of the ethnicity prediction algorithm

RESULTS AND DISCUSSIONS

In total, 153,614 variants were observed in our internal variant DB at a significantly higher allele frequency than in gnomAD. ~77K, ~65K and ~12K variants were found in genes associated with AD, AR and XL diseases, respectively (Table 1). Of these, a total of 6,822 variants were found in ClinVar as P/LP/VUS and manual curation was performed on these variants. As a result, 222, 14 and 3 variants for AD, AR and XL disease genes, respectively, were determined as variants that can be classified as LB with high confidence. Variants that could not be determined as LB were those that were extremely rare internally (<3 heterozygous for AD, <3 homozygous/hemizygous for AR/XL), found in genes reported with incomplete penetrance, variable expressivity, variable onset age or reported as susceptibility/risk genes. Based on this result, we expect that similar percentage of the internal variants not found in ClinVar could also be classified as LB with high confidence even though manual curation of each variant is warranted.

	AD	AR	XL
High frequency inhouse variants (step 3)	77,183	64,551	11,880
P/LP/VUS in ClinVar (Step 4)	1,632	4,905	285
P/LP/VUS ClinVar variants that can be reclassified as LB with high confidence (Step 5)	222	14	3

Table 1. Summary of the variants reclassified

We investigated the ethnicity distribution of the variants we classified as LB to see if underrepresentation of ethnic/population specific variants in gnomAD could have been a factor for not being able to initially classify them as LB. Instead of using the ethnicity information provided by the ordering physicians, we used the predicted ethnicity information as there were many missing data points. The ethnicity of the inhouse samples was predicted by the ethnicity predictor we developed. For the samples that were provided with ethnicity information, the recall rate of the predictor was ~90%. Figure 3a shows the ethnicity distribution of all gnomAD (v2) samples and all 3billion samples. 3billion has significantly larger number of samples predicted as east asians (12,137) even though the total number of 3billion samples is only one seventh of that of gnomAD (20,455/141,456). For the variants reclassified as LB (Step 5), the ethnicity distribution shows even higher proportion of the variants originating from east asian (Figure 3b), suggesting that the main reason these variants could not have been classified as likely benign initially was because they were absent from gnomAD and internal DBs that each ClinVar submitter was using to remove common variants. Additionally, the proportion of 3billion samples predicted as 'others' was also large at ~20% (Figure 3a). This is most likely because there are a lot of (sub)ethnic groups not represented in gnomAD that are present in 3billion DB. Figure 3c shows the country distribution of the ordering institutions.

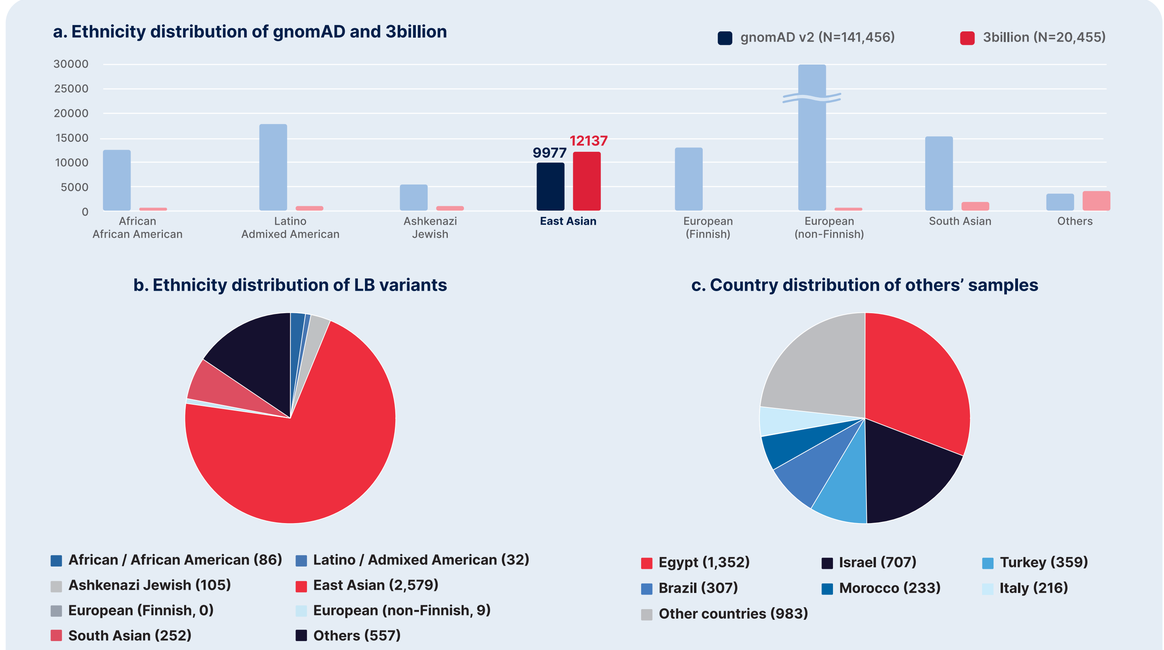

Figure 3. (a) Ethnicity distribution of all gnomAD (v2) samples and all 3billion samples, (b) 3billion samples harboring variants selected in step 5. Country of the ordering institute of 3billion's 'others' is plotted in (c).

Allele frequency information is an essential part of variant classification for diagnosing rare genetic disorders. When a variant that has never been observed in a patient with similar symptoms is identified, and its protein consequence is uncertain, the key factor that determines if the variant is likely pathogenic or likely benign is its frequency in the population. Here we showed that there are a lot of seemingly private variant that could be classified as likely benign because they are commonly found only in certain ethnic group or population. Even though more exome and genome sequencing is being performed globally, there are still underrepresented populations whose data will be crucial in variant interpretation.