ClinVar submitter list leaderboard obscures extensive variation and bias in submission types

Eerik Kaseniit MEng[1], Konrad Karczewski PhD[2], Imran Haque PhD[3]

1. Counsyl Inc. | 2. Massachusetts General Hospital and the Broad Institute | 3. Previously Counsyl Inc., currently Freenome

South San Francisco, California

Introduction

The need for clear and consistent data sharing has been highlighted by recent initiatives like the Cancer Moonshot[1] or tragedies that have led to lawsuits[2,3]. However, contributing variants to ClinVar can be susceptible to the tragedy of the anticommons, as there is cost associated with sharing results, e.g. the person hours required to prepare a submission, with little cost associated with not sharing data. The costs can be reduced by making submissions easy and mitigated by benefits such as elevated status on the ClinVar submitter list[4], ordered by the total number of submissions. Such ranking-based incentive structures could potentially be gamed, e.g. by submitting variants of little "value". We investigated the submission profiles of different submitters in order to understand their competencies and potential biases.

Methods

The regions of interest hypothetically sequenced by a ClinVar submitter were determined by considering the exons of all genes for which there was at least one variant provided by the submitter. Variants in the regions, their allele frequencies (**AF**) and molecular classes (e.g. nonsense mutation) were obtained from gnomAD[5], correcting the allele frequency for rare variants (AF < 0.01)[6]. The 29 submitters submitting at least 1,000 entries in the February 2017 release of ClinVar were each categorized as a clinical diagnostic lab (**CL**), academic lab (**AL**), condition-specific consortium (**CO**), or general database (**DB**). Submitters were hierarchically clustered according to their submission profile defined as the proportion of variants belonging to each combination of clinical significance and AF bin.

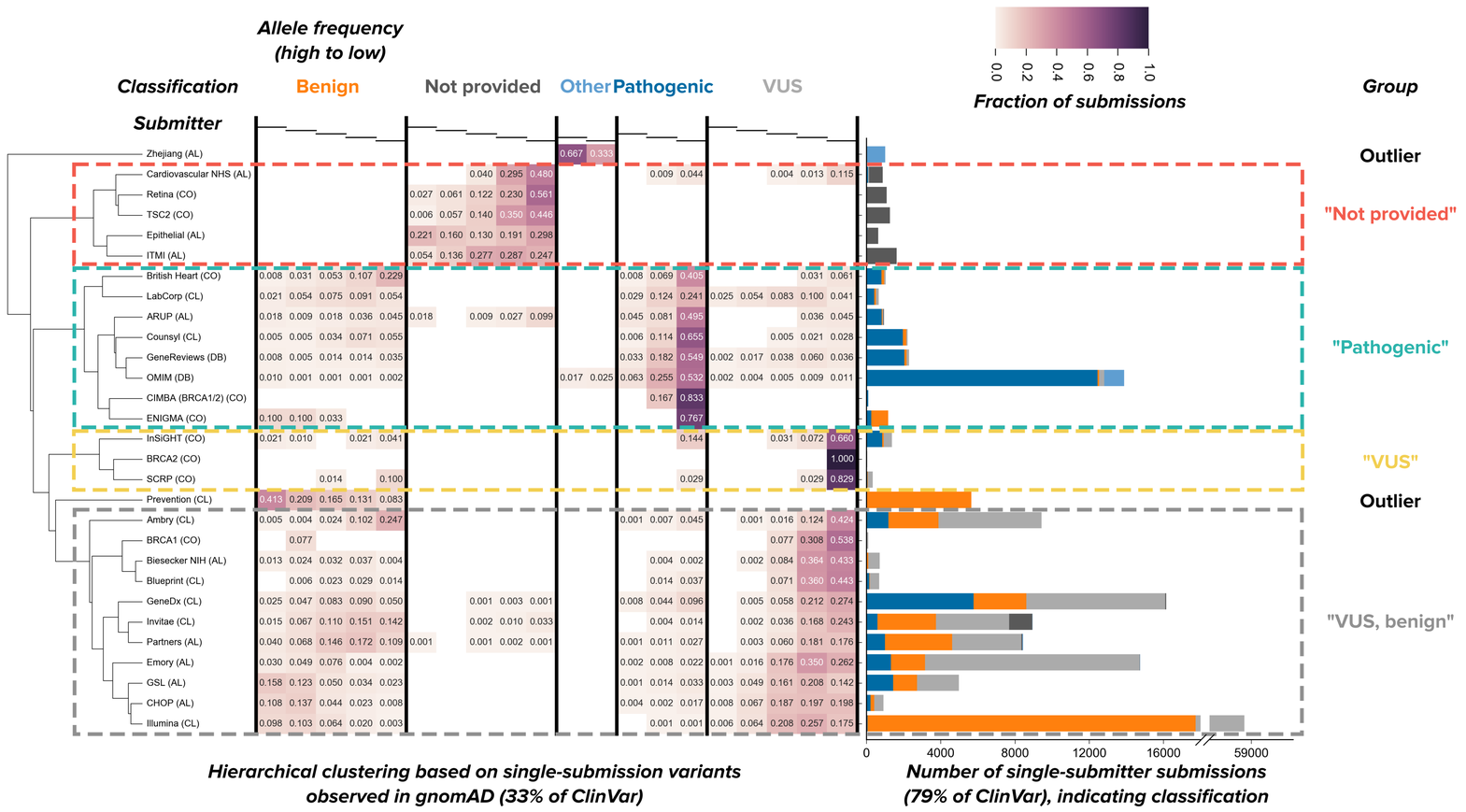
Conclusion

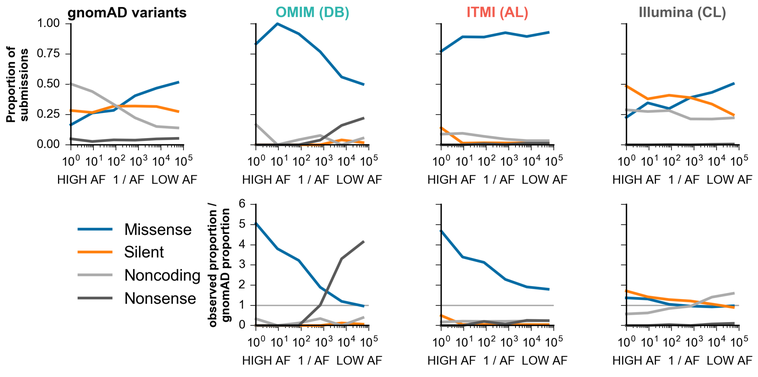
Submitters to ClinVar submit different types of variants when considering their clinical classification and population frequency. Submitters also submit different molecular types of variants. These differential interests in types of variants suggests different motivations of the submitters, so alternative incentives to a general leaderboard tabulated by counting the number of submissions may add value to submitters and ClinVar users, and motivate more submissions and data sharing.

Biases in what kinds of variants and which determinations are submitted could also affect the validity of some submissions, and suggests that some variants are observed, but not interpreted and/or reported on. Since nearly 4 out of 5 variants have a single submitter, these biases should be carefully studied further and the implications considered in clinical practice.

Results

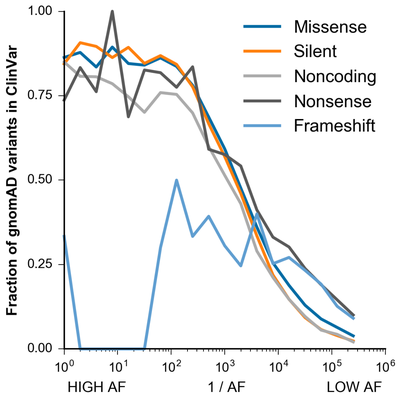
(Right) Hierarchical clustering of submission profiles reveals that organizations submit different types of variant classifications at differing rates. Four main clusters and two outliers (one contributing mainly rare variants with the annotation "cancer" ("Other" here), another contributing mainly common benign variants) emerged in the hierarchical analysis of submission profiles. One cluster was characterized by submitters who mostly provided **pathogenic variants**. Members of another cluster **did not provide any indication to variant pathogenicity** (two consortia, three academic laboratories). Two sets of submitters provided mainly variants of uncertain significance (**VUS**); one set mostly provided **VUS-s** while another provided **VUS and benign variants**. These results do not substantially differ from those obtained from the September 2016 ClinVar dataset, indicating consistency over time. Columns where the maximum value was 2% and cells with value less than 0.05% not shown for display purposes.



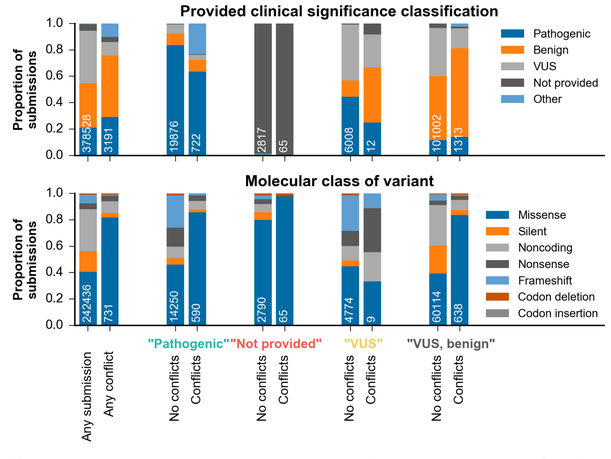


(Above) Submitters show biases in the biological nature of variants for which they submitted classifications. These biases are evident when comparing the proportion of submissions of a particular molecular class within an allele frequency bin to the distribution of variants in gnomAD with the same features. Deviation from the top leftmost pane is quantified in the bottom row for each submitter. Analysis was performed on the common set of 90 genes among these submitters.

REFERENCES: 1. Cancer Moonshot Milestones. https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/milestones | **2.** Willams v Quest/Athena https://www.genomeweb.com/molecular-diagnostics/mothers-negligence-suit-against-quests-athena-could-broadly-impact-genetic | **3.** AMP discussion https://www.genomeweb.com/molecular-diagnostics/amp-meeting-pathologists-discuss-challenges-implementing-acmg-variant | **4.** ClinVar submitter list https://www.ncbi.nlm.nih.gov/clinvar/docs/submitter_list/ | **5.** Lek et al., Analysis of protein-coding genetic variation in 60,706 humans, Nature 2016 | **6.** Whiffin et al., 2016 Using high-resolution variant frequencies to empower clinical genome interpretation, doi: https://doi.org/10.1101/073114



(Above) Rare variation remains uncharacterized in ClinVar. Most of the variation that is prevalent at an allele frequency of about 1 in 1000 (or more common) for noncoding variants, 1 in 2000 (or more common) for missense or silent variants, and 1 in 4000 (or more common) for nonsense variants is captured, but frameshift variants are poorly represented. Genes with at least 5 submitters were included in the analysis.



(Above) Biases are also observed in the case of conflicting submissions. Conflicts tend to involve more benign and pathogenic mutations; there are fewer VUS determinations in the case of a conflict than overall (compare "Any submission" & "Any conflict" plots, top panel). Conflicts are generally enriched for missense mutations, while conflicts for silent, noncoding, or frameshift mutations are less common (compare "Any submission" & "Any conflict" plots, bottom panel). Biases for the submitter clusters are evident when comparing the "Any submission", "Any conflict", and "conflicts" plots. A classification of "Other" was considered conflicting; a VUS classification was not.

View all posters and research at
research.counsyl.com