Mammalian species conservation data and the implications for clinical variant classification

Yuya Kobayashi*, Aurora Alvarez-Buylla*, Paige Taylor*, Shan Yang, Keith Nykamp, and Scott Topper

Invitae, San Francisco, CA

*These authors contributed equally.

Introduction

Functionally critical amino acid residues are subject to evolutionary pressure and are expected to be conserved, whereas non-critical positions may be more tolerant to variation. As such, evolutionary conservation data may be a powerful component of predicting the deleterious effect of missense mutations. In practice, however, the predictive power of existing conservation-based in silico algorithms has been inadequate for broad clinical application; the 2015 ACMG guidelines for the interpretation of sequence variants consider consensus among multiple algorithms to be only supporting evidence for variant classification. However, the guidelines also discuss an alternative application for conservation data: the presence of the variant amino acid change in multiple nonhuman mammalian species is designated strong evidence for a classification of benign. This distinction reflects the assumption that variation within the mammalian clade speaks more directly to questions related to mammalian physiology and is therefore more relevant to the question of human disease.

To assess the predictive power of this type of data, we analyzed 43,387 missense variants from ClinVar, binned the data by classification, and examined conservation across 61 nonhuman mammalian species.

ACMG Guidelines

On computational (in silico) prediction algorithms:

"Not overestimating computational evidence is important. . . . [M]ost algorithms have not been validated against well-established pathogenic variants." Therefore, these predictions are considered only supporting evidence.

On mammalian conservation data:

"The variant amino acid change being present in multiple nonhuman mammalian species in an otherwise well-conserved region, suggesting the amino acid change would not compromise function, can be considered strong evidence for a benign interpretation."

E.g., [*TP53*] p.Arg209Lys: Lysine is observed at this codon in seven mammalian species (among monkeys, hamsters, and bats). An arginine-to-lysine change at this codon is likely to be tolerated in humans as well.

Data Sources

Interpreted variants:

ClinVar: Filtered to include only interpretations from CLIA-certified laboratories: 43,387 interpreted missense variants.

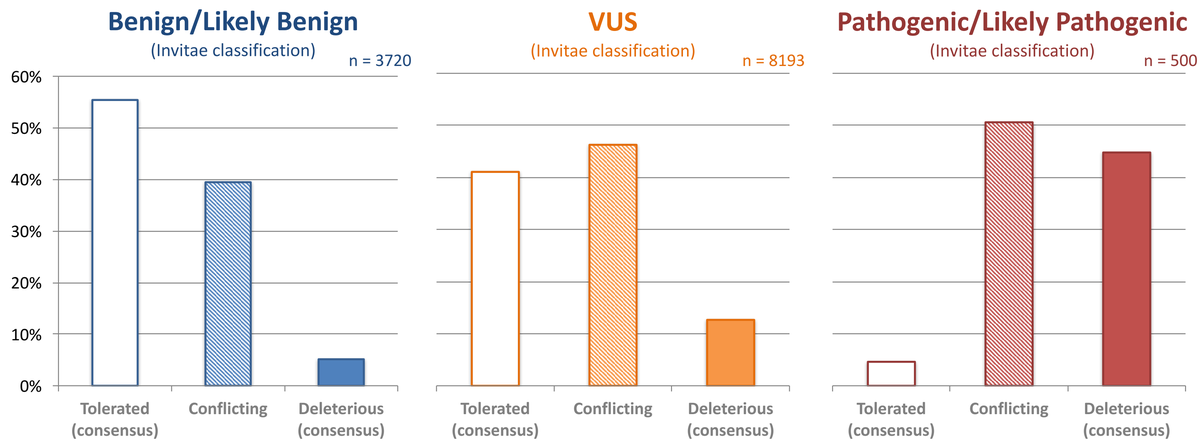
Computational (in silico) predictions:

Pathogenicity predictions were made by using **SIFT**, **Polyphen-2**, and **AlignGVGD** with default settings.

Mammalian conservation data:

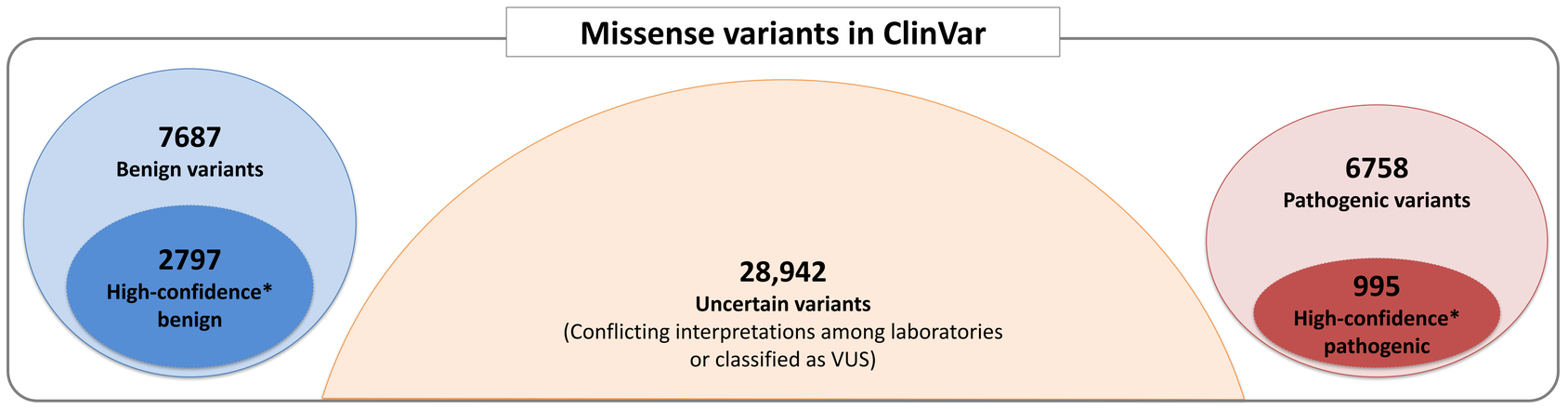
Amino acid conservation data were derived from the **Vertebrate Multiz Alignment and Conservation** data set in the UCSC Genome Browser. Alignments were restricted to the 61 nonhuman mammalian species.

Accuracy of Computational Prediction Algorithms



1. Prediction algorithms conflicted with each other **44%** of the time (5512 of 12,413 variants).
2. The consensus prediction matched actual interpretation only **54%** of the time (2285 of 4220 non-VUS variants).
1. **92%** of variants that are predicted deleterious by SIFT or Polyphen2 are benign, likely benign, or VUS (data not shown).

Predictive Power of Mammalian Conservation Data



*A variant classification was designated "high-confidence" if (1) there was a consensus among multiple laboratories, or (2) the variant was assessed by Invitae.

Presence of the variant amino acid in multiple mammalian species, by ClinVar classification

	Benign		Uncertain	Pathogenic	
	All	High confidence	All	All	High confidence
The variant amino acid **not present** in multiple mammalian species (conserved position)	4513	1716	23,876	6691	**987**
The variant amino acid **is present** in multiple mammalian species (non-conserved position)	3174	1081	5066	94	**8**
Ratio	**1.4**	**1.6**	**4.7**	**71.2**	**123.4**

Eight pathogenic variants (high confidence) with the variant amino acid present in multiple species

Gene	DNA change	Protein change	Number of species with variant AA	Notes
MEFV	c.2177T>C	p.Val726Ala	31	• Associated with familial Mediterranean fever. • Relatively common in the European population. • May be **undergoing positive/balancing selection owing to heterozygote advantage** (PMID: 11242116).
MEFV	c.2040G>C	p.Met680Ile	11	
MEFV	c.2230G>T	p.Ala744Ser	4	
MEFV	c.2084A>G	p.Lys695Arg	3	
TTR	c.424G>A	p.Val142Ile	5	• Associated with late-onset hereditary amyloidosis. • Certain species of **monkeys** are known to carry this variant and also **develop late-onset amyloidosis** (PMID: 22184092).
BRCA1	c.5453A>G	p.Asp1818Gly	2	• Shown to **disrupt mRNA splicing** (PMID: 20875879).
PRSS1	c.47C>T	p.Ala16Val	2	• Suggested to "contribute to multigenic inheritance of a **predisposition to pancreatitis**" (PMID: 19951905). • Very common (5%) in the African population.
SLC26A4	c.85G>C	p.Glu29Gln	2	• Associated with Pendred syndrome.

Conclusions

1. **Computational protein effect prediction algorithms are highly inaccurate and of limited utility in a clinical setting.**

2. **The presence of the variant amino acid in multiple mammalian species has strong predictive power in identifying variants that are not pathogenic.**
   - This method has a false prediction rate of less than 1%. These false predictions appear to occur for biological, not technical, reasons.
   - Many benign variants are also conserved and therefore should not be used to predict for variant that are pathogenic