Published in final edited form as:

Nat Methods. 2012 September ; 9(9): . doi:10.1038/nmeth.2147.

Evolutionary Diagnosis Method for Variants in Personal Exomes

Sudhir Kumar^{1,2}, Maxwell Sanderford¹, Vanessa E. Gray¹, Jieping Ye^{1,3}, and Li Liu¹

Sudhir Kumar: s.kumar@asu.edu

¹Center for Evolutionary Medicine and Informatics, Biodesign Institute, Arizona State University, Tempe, Arizona, USA

²School of Life Sciences, Arizona State University, Tempe, Arizona, USA

³School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, Arizona, USA

To the Editor

Each human exome contains thousands of nonsynonymous single nucleotide variants (nSNVs) of unknown effect, whose potential impact on biological function is now routinely assessed using computational methods for use in biomedical research and in clinical genome profiling reports. Of the variants receiving a non-neutral (function-damaging) prediction, those at evolutionarily conserved sites are frequently of heightened interest for scientists and clinicians, because these sites are among the most critical for proper protein function. Indeed, a majority of amino acid mutations that have been investigated experimentally are located at ultra-conserved sites, which show no amino acid residue difference in diverse species spanning over five hundred million years of evolution (Supplementary Fig. 1).¹ Functionally damaging mutants at these sites are likely to have significant consequences in health and disease.

For these ultra-conserved sites, we estimated the false positive rate (FPR) of Condel² and PolyPhen-2³ by using a standard collection of neutral variants (HumVar³) that was used to train and test these two state-of-the-art tools. Our analysis revealed a very high FPR for Condel (89%) and PolyPhen-2 (75%). For 73% of the neutral nSNVs in HumVar, both tools produced a damaging prediction. The overall accuracy of these tools at ultra-conserved positions was also low (55% and 60%, respectively). Therefore, predictions produced by current computational tools may mislead downstream experimental and clinical investigations aimed at studying functionally important sites.

To remedy this problem, we have developed a method that adaptively fits the best prediction model for nSNV sites evolving at distinctly different rates, as revealed by multispecies comparison (ultra-, well- and less-conserved, Supplementary Fig. 1). Our Evolutionary Diagnosis (EvoD) method considers fundamental evolutionary properties of affected site and amino acid changes, including measures of the biochemical severity and evolutionary probability of the amino acid change, as well as multi-level taxonomic evolutionary rates and timespans of the position affected. Within a sparse-learning framework, the ability of these properties to explain the observed (training) data was evaluated in EvoD for establishing an efficient predictive statistical model (Supplementary Methods). A web server for evaluating novel variants using EvoD is available at <http://barn.asu.edu/EvoD>.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Note: Supplementary information is available on the Nature Methods website.

At ultra-conserved sites, EvoD led to large reductions in the FPR: 55% for Condel and 41% for PolyPhen-2. We retrieved the population allele frequency of the neutral HumVar nSNVs at ultra-conserved sites from a 5,400 exome dataset⁴ and found that EvoD improved diagnoses across the spectrum of rare (<0.1%) to common (>5%) alleles (Fig. 1a). The balanced accuracy of EvoD was also significantly higher at ultra-conserved sites (Table 1; $P < 10^{-10}$). Furthermore, EvoD's performance was consistent across ultra-, well-, and less-conserved sites, whereas Condel and PolyPhen-2 showed uneven performance across these classes (Table 1).

For each nSNV, the EvoD statistical model also produced an impact score that reflected the degree of neutrality: most neutral = 0 and most non-neutral = 100. In an analysis of 244,272 nSNVs from the 1000 Genomes Project, we found that the population frequency of nSNPs decayed with increasing impact score (Fig. 1b). Therefore, we used the empirical distribution of EvoD scores to determine the statistical significance of a neutral or non-neutral diagnosis (P -value) adaptively for variants at ultra-, well-, and less-conserved sites (Supplementary Fig. 2).

Using EvoD, we analyzed nSNVs in an example set of eight personal HapMap exomes⁵, which contains a total of 13,372 nSNVs at ultra-conserved sites. 4% of these were predicted to be non-neutral ($P < 0.05$). An overwhelming majority (94%) of these non-neutral nSNVs were found in heterozygous genotypes that would neutralize the negative effects of recessive alleles. Similar results were observed for 35,367 nSNVs at well- and less-conserved sites, which were also reflected in the neutrality heat maps showing the EvoD impact scores of nSNVs in heterozygous genotypes (Fig. 1c). In contrast, the fraction of homozygous nSNVs with high EvoD impact scores is much smaller at ultra-conserved sites (Fig. 1c). EvoD predicted no more than one homozygous nSNV per exome to be non-neutral ($P < 0.05$) in ultra-conserved sites, which is consistent with the fact that individuals contributing to HapMap sequencing do not suffer from any known Mendelian disease.

Our results show that an evolution-aware approach to training and testing computational tools leads to better functional predictions for nSNVs, particularly at the most functionally important positions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Nevin Gerek, Crystal Hepp, and Mia Champion for many insightful comments, and Joshua Akey for providing the 5,400 exomes data. Ms. Carol Williams provided editorial support. This research was supported by a research grant from the National Library of Medicine (R01 LM010834) and a Postbaccalaureate Research Education Program grant (R25GM071798) from National Institutes of Health.

References

1. Gray VE, Kukurba KR, Kumar S. *Bioinformatics*. 2012; 33610.1093/bioinformatics/bts
2. Gonzalez-Perez A, Lopez-Bigas N. *Am J Hum Genet*. 2011; 88:440–449. [PubMed: 21457909]
3. Schmidt S, et al. *Nat Methods*. 2010; 7:248–249. [PubMed: 20354512]
4. Tennessen JA, et al. *Science*. 2012; 337:64–69. [PubMed: 22604720]
5. Ng SB, et al. *Nature*. 2009; 461:272–276. [PubMed: 19684571]
6. Matthews BW. *Biochim Biophys Acta*. 1975; 405:442–451. [PubMed: 1180967]

Figure 1.

Performance and application of the new EvoD method. **(A)** Improvement in reducing the False Positive Rate, FPR, for nSNVs occurring with different population frequencies at ultra-conserved sites (Condel: black bars; PolyPhen-2: white bars). Allele frequency data for HumVar2 neutral nSNVs were retrieved from the 5,400 exome dataset⁴. **(B)** The relationship of EvoD impact scores and the population frequencies of 244,272 nSNVs from the 1000 Genomes Project. Each point shows the average allele frequency for nSNVs with impact scores in increments of 5. **(C)** Neutrality heat maps based on EvoD predictions for homozygous (Hom.) and heterozygous (Het.) nSNVs from eight HapMap exomes that occur at ultra-, well- and less-conserved sites. Heat maps were constructed by sorting nSNVs by impact score and assigning colors from dark blue (most neutral) to red (most non-neutral) on a linear scale according to the estimated *P*-value.

Table 1
Performance of EvoD, Condol and PolyPhen-2 (PP2) evaluated using the HumVar2 dataset.

Method	Evol. Consv.	HumVar2 Neutral			HumVar2 Disease-associated			Diagnosis Rate			Accuracy	
		True Negative	False Positive	False Negative	True Positive	TNR	FPR	FNR	TPR	BAcc	MCC	
EvoD	Ultra	604	317	2,528	7,045	66%	34%	26%	74%	70%	39%	
	Well	2,622	1,247	1,360	4,636	68%	32%	23%	77%	73%	45%	
	Less	6,883	3,382	4,98	1,409	67%	33%	26%	74%	70%	41%	
Condel	Ultra	103	818	111	9,462	11%	89%	1%	99%	55%	21%	
	Well	1,500	2,369	393	5,602	39%	61%	7%	93%	66%	38%	
	Less	6,183	4,082	588	1,319	60%	40%	31%	69%	65%	30%	
PolyPhen-2	Ultra	226	664	556	8,552	25%	75%	6%	94%	60%	26%	
	Well	2,585	1,356	1,190	4,589	66%	34%	21%	79%	73%	45%	
	Less	9,398	898	1,177	625	91%	9%	65%	35%	63%	31%	

TPR: True Positive Rate (sensitivity), FPR: False Positive Rate, TNR: True Negative Rate (specificity), FNR: False Negative Rate, Bacc: Balanced Accuracy; MCC: Matthews Correlation Coefficient⁶.
Bacc = ½(TPR + TNR)/2; MCC = (TPR×TNR – FPR×FNR) / ((TPR+FPR)×(TNR+FNR)×(TNR+FN|R)).