Genomic variant annotation using semi-supervised clustering outperforms existing consensus and variant-type-specific methods

Counsyl

Sharad Vikram, Matthew D. Rasmussen PhD, Eric A. Evans PhD, Imran S. Haque PhD

South San Francisco, CA

Introduction

The assignment of pathogenic or benign status to variants observed in an individual ("variant interpretation") has increasingly become the primary bottleneck in the clinical analysis of next-generation sequencing data.

We describe SSCM ("Semi-Supervised Clustering of Mutations"), a fully-probabilistic ensemble approach to variant interpretation. SSCM is available at **https://github.com/counsyl/sscm**.

Results

We compared SSCM's performance at predicting pathogenic variants to that of popular variant effect predictors: CADD, SIFT, PolyPhen2, verPhyloP, and verPhastCons for missense variants; CADD, HSF, NNSplice, and MaxEnt for noncanonical splicing variants.

Additionally, to test SSCM's ability to separate pathogenic from damaging variants, we test it on pathogenic loss-of-function (LoF) variants against benign LoF variants[2].

Methods

SSCM uses Bayesian semi-supervised mixture learning, a machine learning technique that overcomes sparse training data and missing input features (e.g. those only defined on parts of the genome). Like CADD[1], it uses simulated variants as training input, but models them as a mixture of benign and pathogenic variants to learn a benign-vs-pathogenic classifier, rather than a benign-vs-simulated classifier (Figure 1).

Conclusion

SSCM shows superior performance compared to existing ensemble methods on a variety of difficult classes of SNVs, including missense, noncanonical splicing, and benign loss-of-function variation.

All Counsyl posters available online at research.counsyl.com

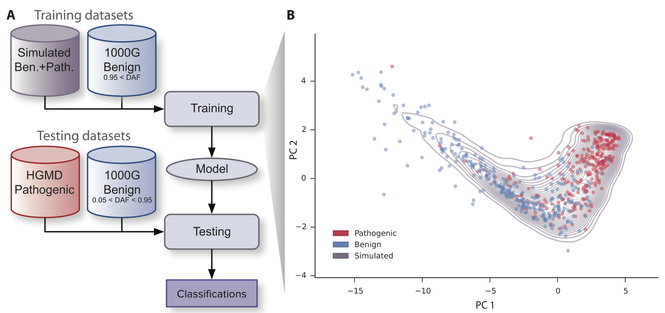
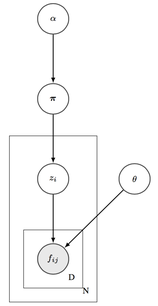


Figure 1

a) We trained SSCM following CADD's procedure, using 95% DAF variants from 1000 Genomes as a benign truth set and simulated variants as an unlabeled mixture of benign and pathogenic. Pathogenic variants from HGMD and ClinVar were used to evaluate the classifier, as well as known loss-of-function tolerant variants[2]. **b)** PCA of simulated variants (gray) against known benign (blue) and pathogenic (red) shows the mixed nature of simulated variants.



Feature name	Coverage
verPhyloP	Genome-wide
verPhastCons	
Gerp++RS	
ENCODE H3K27Ac	
ENCODE H3K4Me1	
ENCODE H3K4Me3	
SIFT	Missense
PolyPhen2	
HSF	Splice sites
MaxEnt	
NNSplice	

Figure 2

a) SSCM's model can be described by a Bayesian network. Conditioned on its assigned cluster z_i (e.g. benign, deleterious), a variant i has independent features f_{ij} (e.g. conservation), each with its own multinomial or Gaussian distribution parameterized by θ. Cluster assignments are modeled with a multinomial prior of mixing weights π, which in turn has a Dirichlet prior with hyperparameter α.
b) List of features used in final SSCM model and their domains of applicability.

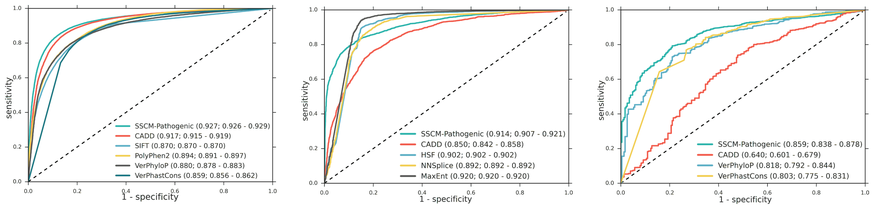


Figure 3: ROC AUC for benign vs pathogenic variant prediction

Left to right: **a)** Missense variants from HGMD. **b)** Noncanonical splicing variants from HGMD. **c)** LoF-tolerant variants vs HGMD pathogenic LoF variants. SSCM is able to significantly outperform CADD because of its better integration of conservation scores.

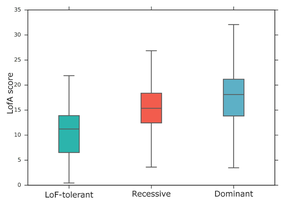


Figure 4: Gene effect prediction

Left: The average SSCM score for all stop-gain variants in a gene ("LoFA") is significantly different depending on whether LoF of the gene is recessive-effect, dominant-effect, or the gene is LoF-tolerant ($p < 1e-7$ for all pairs).

CITATIONS 1| Kircher M et al. Nat Genet 2014 Mar;46(3):310-5 2| MacArthur DG et al. Science 2012 Feb 17;335(6070):823-8