Challenges in variant annotation for clinical genomic testing

1734F

Jennifer Yen, Sarah Garcia, Aldrin Montana, Steve Chervitz, Brian Linebaugh, John West, Richard Chen and Deanna M Church

Personalis, Inc. | 1330 O'Brien Drive, Menlo Park, CA 94025

Contact:
jennifer.yen@personalis.com

Problem

Generating accurate HGVS nomenclature is dependent on successful execution of the following:

a. Identifying the correct transcript version

Due to transcript complexity, it is important to identify the clinically relevant transcript when annotating and reporting a variant.

Up-to-date transcript versions should also be observed, as small changes in versions may impact the coding sequence.

b. Left or right justification of the sequence variant

	Pos. vcf (left shift)	Pos. HGVS (right shift)
Ref ACCTTTTTGTCTG		
Alt ACCTTTTTTGTCTG	4	9

c. Translating the annotation from transcript to protein

NM_003119.3:c.90dupT → NP_003110.1:p.Pro31Serfs*43

Errors in any of these steps can lead to ambiguous and/or incorrect HGVS representation.

Methods

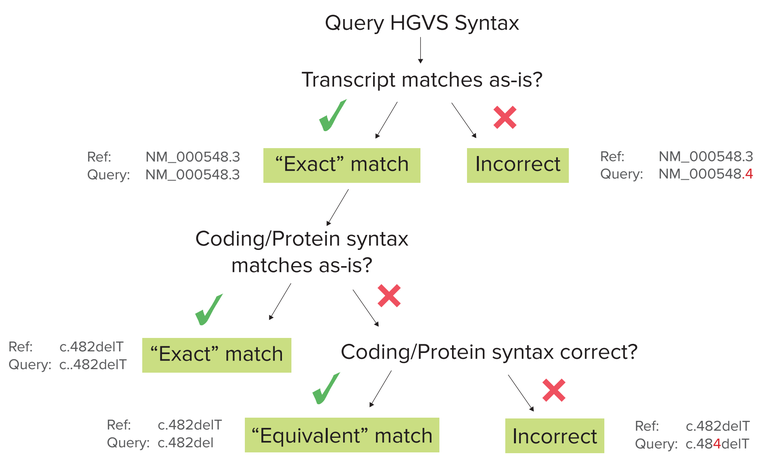
We tested three tools:

Table 1. Tools Used

Tool	Speed (100K variants)	Implementation
SnpEff[1]	35 min	Easy
VEP[2]	3 hours	Easy
Variation Reporter[3]	4 days	Difficult

** Mutalyzer was not assessed as the tool was used to determine the reference syntax in the test set.*

Figure 1. Method for Assessing Syntax Accuracy



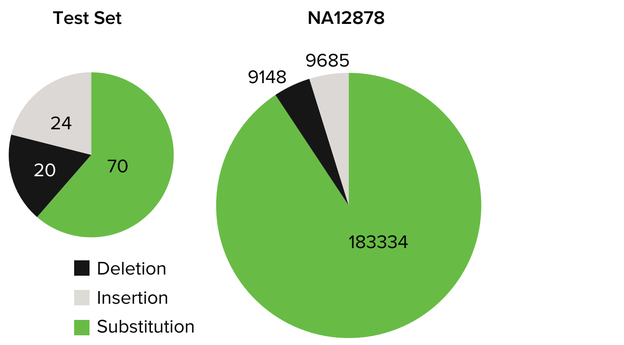
Results

A standard 'truth' set for evaluating HGVS syntax

To evaluate the accuracy of tools for generating HGVS nomenclature, we created a 'truth' dataset of 115 variants across numerous variant types and effect impacts, and manually curated their HGVS syntax.

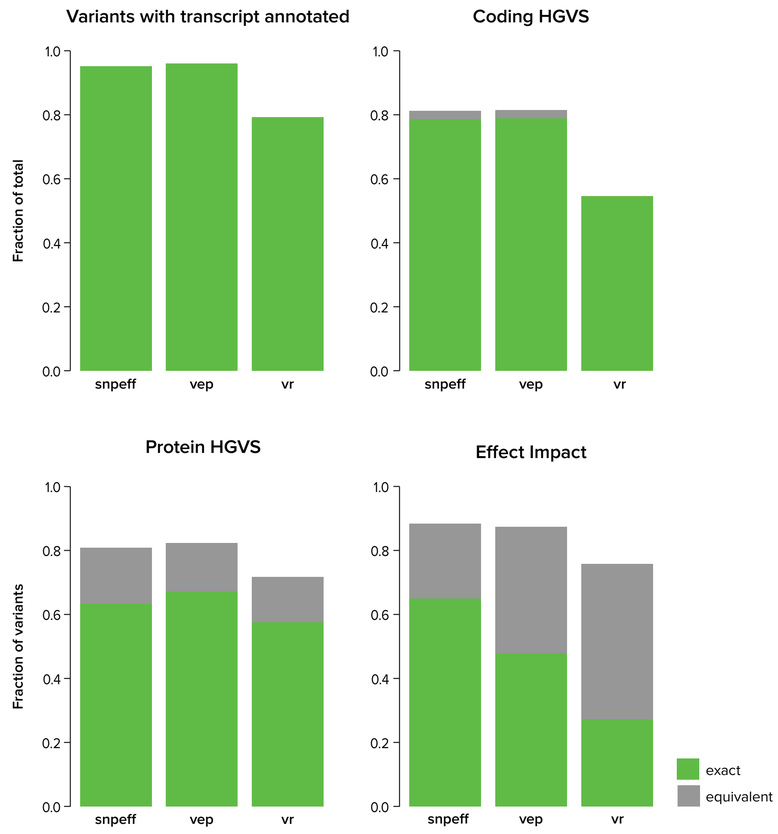
To deeply evaluate the robustness of these tools, we included variants that would be particularly difficult to annotate. For example, insertions and deletions have greater representation as a proportion in our test set compared to both the ClinVar dataset and the exome of CEPH NA12878.

Figure 2. Test Set Contents by Variant Type



We tested how well the tools could navigate the current reference assembly by including variants in complex regions, such as in sequences with alternative representation to the reference (novel patches) or scaffold sequences that have been updated from GRCh37 (fix patches).

Figure 3. Performance of Tools on Test Set



Overall, we found that SnpEff and VEP have comparable output, while Variation Reporter performed far worse in at generating both coding and protein HGVS syntax. Neither tool accurately annotated all variants across variant types correctly.

ClinVar Dataset

We next assessed the concordance between the tools with nomenclature in the ClinVar dataset[4], which reflects variants that would be reported in a clinical setting. We extracted the HGVS syntax and effect impacts of 113K variants using a modified version of a parsing script from the MacArthur Lab[5,6].

Figure 4. ClinVar Contents by Variant Type

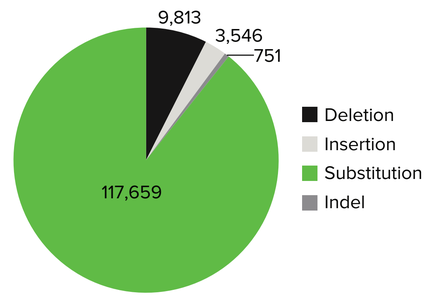
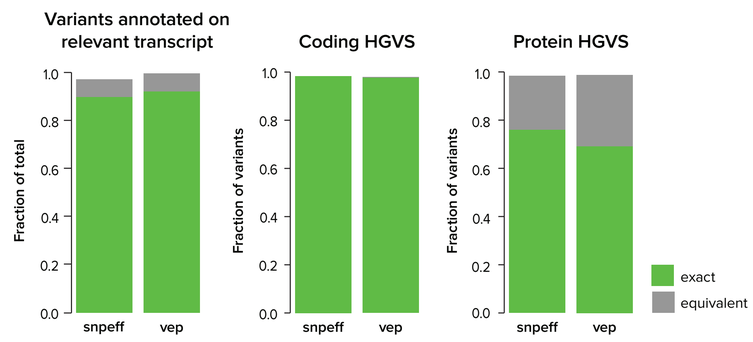


Figure 5a. Overall Concordance of Tools with ClinVar Dataset



- Approximately 7.5% of variants were not assessed because of differences in transcript versions (grey).
- Although both tools performed well at annotating SNVs, neither annotated non-SNV types with 100% accuracy.

Figure 5b. Concordance of Tools with ClinVar by Variant Type

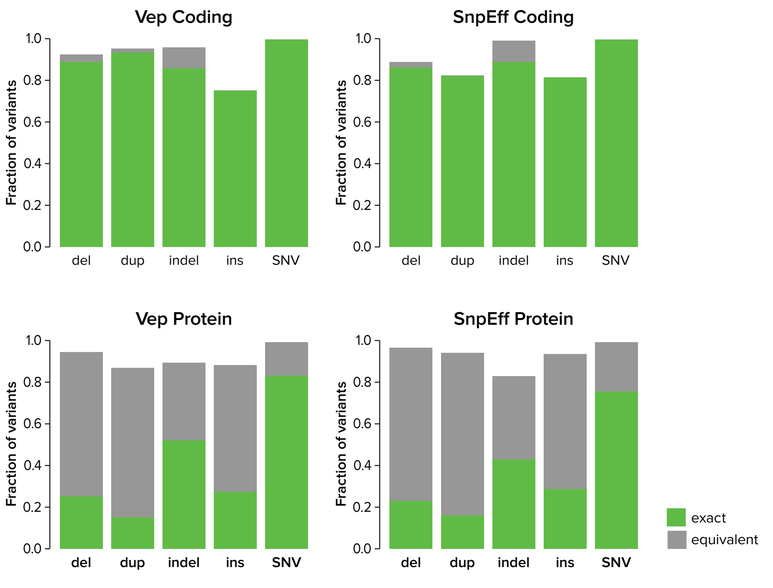


Table 2a. Coding Syntax Synonyms

coding variant type	ClinVar	SnpEff	Vep	VR	Reference ID
duplication	c.567_568dup	c.567_568dupTT	c.567_568dupTT	c.567_568dupTT	rs137854332
deletion	c.562_563delCA	c.562_563delCA	c.562_563delCA	c.564delGinsCAG	PTV003
indel	c.711_734delinsCCCC	c.711_734delTGAGAGCCGGCT-GGCGGATGCGCTinsCCCC	c.711_734delTGAGAGCCGGCT-GGCGGATGCGCTinsCCCC	c.711delTinsCCCC	PTV004

Table 2b. Protein Syntax Synonyms

protein effect	ClinVar	SnpEff	Vep	VR	Reference ID
frameshift_variant	p.Asn846Ilefs	p.Asn846fs	p.Asn846IlefsTer13	p.Asn846Ilefs	rs386134183
frameshift_variant	-	p.Glu238fs	p.Glu238ProfsTer9	p.Phe237_Glu238insPro	PTV004
inframe_insertion	-	p.Arg309_Arg310insArgArg	p.Arg310_Arg311dup	p.Arg311_Lys312insArgArg	PTV085
inframe_deletion	-	p.Ala1111_Ala1119del	p.Ala1111_Ala1119del	p.Ala1119_Gly1120insAlaAlaAlaAlaA-laAlaAlaAlaAla	PTV106
stop_gained	p.Gln100Ter	p.Gln100*	p.Gln100Ter	p.Gln100Ter	rs119103276
synonymous_variant	p.Arg317=	p.Arg317Arg	p.=	p.Arg317=	rs111033272

Conclusions

- The correct transcript identification is important for HGVS syntax assessment at the protein and coding level.
- There is significant variability in nomenclature used to describe variants across tools and datasets.
- Non-SNV syntax is not always correct: should be reviewed before clinical reporting.
- To avoid ambiguity, variants should always be reported by their genomic coordinates.

References

1. Cingolani P, Platts A, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly.* 2012 Apr–Jun;6(2):80–92.
2. McLaren W, Pritchard B, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics.* 2010 Aug 15;26(16):2069–70.
3. Variation::Reporter - A perl module to access NCBI Variation Reporter service. [API] (2015). Retrieved from http://www.ncbi.nlm.nih.gov/variation/tools/reporter/docs/api/perl.
4. Landrum MJ, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014 Jan 1;42(1):D980–5.
5. ClinVar Dataset [XML]. (July 2015). Retrieved from FTP site: ftp.ncbi.nlm.nihgov.
6. Minikel E and MacArthur, D. Parsing ClinVar data. [GitHub Repository] (2015). https://github.com/macarthur-lab/clinvar

Personalis®
Pioneering Genome-Guided Medicine