Challenges in variant search and annotation for clinical cancer testing

Jennifer Yen, Sarah Garcia, Aldrin Montana, Steve Chervitz, John West, Richard Chen and Deanna M Church
Personalis, Inc. | 1330 O'Brien Dr, Menlo Park, CA 94025

Contact:
jennifer.yen@personalis.com

Problem

Clinical genomic testing requires the robust generation and identification of variant-level information in relation to disease.

Standard transcript and protein variant nomenclature is based on guidelines by the Human Genome Variation Society (HGVS).

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

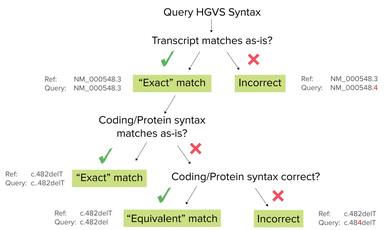
We tested three tools:

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



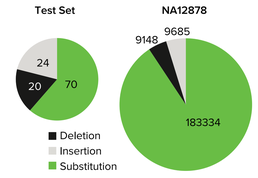
Results

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.

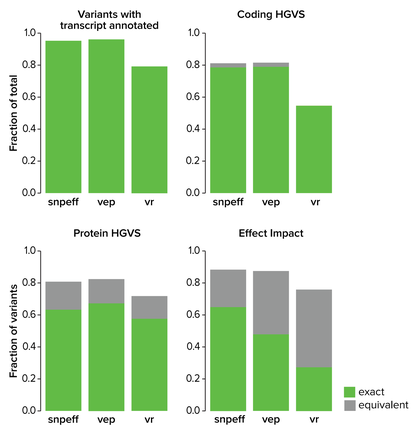
To deeply evaluate the robustness of these tools, we included variants that would be particularly difficult to annotate.

Figure 2. Test Set Contents by Variant Type



We included a greater proportion of insertions and deletions, and 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

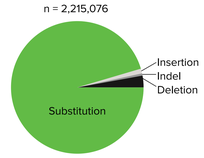


Overall, we found that SnpEff and VEP have comparable output, while Variation Reporter performed far worse in at generating both coding and protein HGVS syntax. Neither tool accurately annotated all variants across variant types correctly.

Syntax concordance with COSMIC

We next assessed the concordance between the tools with nomenclature in the COSMIC dataset[4], a common resource for clinical cancer labs. Because COSMIC employs a different transcript collection (Ensembl) from our clinical lab (NCBI), we performed a transcript-independent assessment. That is, we sought to determine if there was any concordance between the tools and database.

Figure 4. COSMIC Small Variants by Type



- The tools recapitulated only 5–75% of variant nomenclature present in COSMIC. Concordance of equivalent syntax was below 90% except for VEP annotation of SNVs.
- In contrast to VEP and SnpEff, COSMIC reports all duplications as insertions.
- Even though COSMIC employs VEP for annotation, the concordance between VEP and COSMIC still varies, particularly for coding HGVS.

Figure 5. Concordance with COSMIC by Variant Type

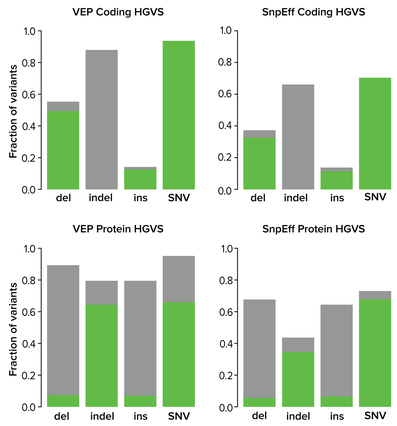


Table 2a. Coding Syntax Synonyms

type	position	ref	alt	COSMIC	SnpEff	Vep	Reference ID
indel	chr1:109446761	TGGT-GAACTAA-CAGCAC	GCTTTGA	c.1077_1093>GCTTTGA	c.1077_1093delTGGT-GAACTAACAGCACins-GCTTTGA	c.1077_1093delTGGT-GAACTAACAGCACins-GCTTTGA	COSM5194180
duplication	chr12:4665573	C	CA	c.422_423insA	c.428dupA	c.428dupA	COSM4719972
deletion	chr1:20976972	TG	T	c.1535delG	c.1538delG	c.1538delG	COSM5129962

b. Protein Syntax Synonyms

effect	position	ref	alt	COSMIC	SnpEff	Vep	Reference ID
duplication	chr1:27023017	G	GGCA	p.Ala45_Glu46insAla	p.Ala42dup	p.Ala45dup	COSM3732828
termination	chr1:26752948	G	A	p.*210*	p.Ter210Ter	p.=	COSM3487544
3' extension	chr1:26603784	A	C	p.*763Cys	p.Ter763Cysext*?	p.Ter763CysextTer14	COSM324933
synonymous	chr12:14975926	C	T	p.Phe19Phe	p.Phe19Phe	p.=	COSM1299184
indel	chr12:2062339T	T	TCGC	p.Gln256>ArgGlu	p.Gln256delinsArgGlu	p.Gln256delinsArgGlu	COSM1741200
frameshift	chr12:25356904	ATTCT	A	p.Ser6fs*1	p.Ser6fs	p.Ser6Ter	COSM1476431
frameshift	chr1:22965350	TC	T	p.Gln64fs*>182	p.Gln64fs	p.Gln64LysfsTer218	COSM4667206

Conclusions

- There is significant variability in variant nomenclature among annotation tools and COSMIC.
- Non-SNV syntax is not always accurate and should be reviewed prior to clinical reporting.
- To minimize ambiguity, variants should be primarily referenced 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. Forbes SA, Beare D, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015 Jan;43:D805-11.

Personalis®
Pioneering Genome-Guided Medicine