Challenges in variant annotation for clinical genomic testing

1734F

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

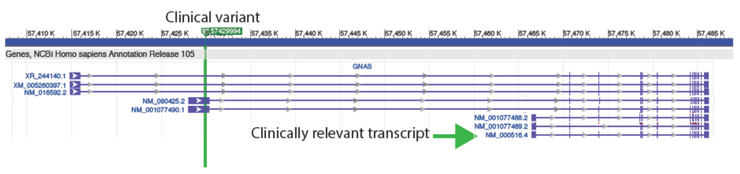
Personalis, Inc. | 1330 O'Brien Drive, Menlo Park, CA 94025

Contact:
jennifer.yen@personalis.com

Problem Overview

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

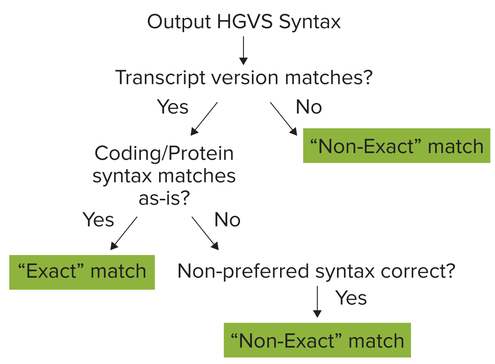
We tested three tools:

Table 1. Tools Used

Tool		Source
SnpEff[1]		Pablo Cingolani
Variation Reporter[2]	(vr)	NCBI
Variation Effect Predictor[3]	(vep)	Ensembl

* 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



Summary

Table 2. Tool Summary

Tool	Speed (100K variants)	Difficulty of implementation
SnpEff	Fast – 10 minutes	Easy
VEP	Medium – 8 hours	Easy
Variation Reporter	Slow – 4 days	Difficult

Conclusion

- The correct identication of the transcript and version has a significant impact on the HGVS syntax assessment at the protein and coding level.
- Evaluation on a standard 'truth' is important for defining the strengths and limitations of these tools.

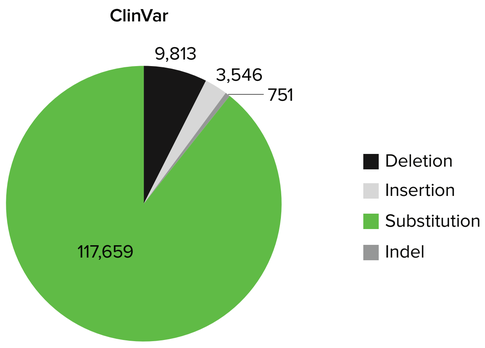
References

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

ClinVar Dataset

To assess the overall accuracy of tools for generating HGVS nomenclature, we tested the tools on 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 2. ClinVar Contents by Variant Type



* Note: Results presented are preliminary findings from our analysis.

Results

Figure 3a. Performance of Tools on ClinVar Dataset

Correct annotation as a fraction of total variants (dependent on identifying the correct transcript)

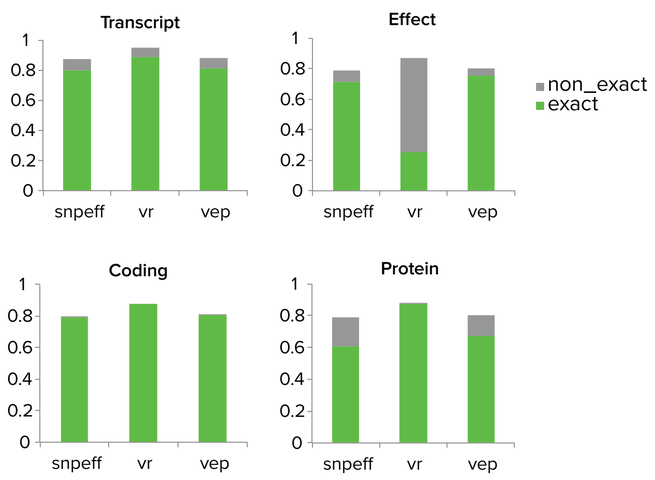
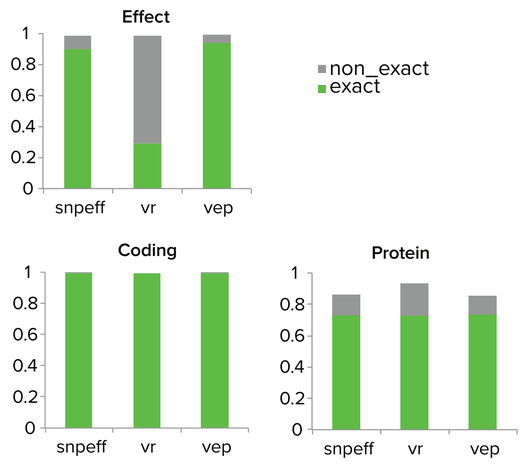


Figure 3b. Performance of Tools on ClinVar Dataset

Correct annotation as a fraction of variants attempted (independent of transcript annotation set)



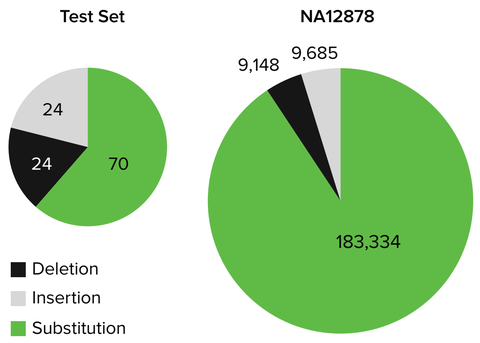
- Variation Reporter identified the most transcripts compared to SnpEff and VEP (defined by GRCh37, 105), and thereby correctly annotated a greater number of variants at the coding and protein level.
- As a fraction of total variants attempted, all three tools performed similarly well, with some deviations from the preferred HGVS syntax.

A standard 'truth' set for evaluating HGVS syntax

The ClinVar dataset is well known for its complex content and data structure. To better control for content, 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 1).

Figure 4. 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) (Table 3).

Table 3. Contents by Genomic Features

Genomic Features	Count
RefSeq/Reference sequence mismatches	7
Paralog Regions	2
Patch regions (novel or fixed)	2
Regions unique to alt	3
Pseudogene with no transcript	1
SHANK2 region, no transcript	1

Results

Figure 5a. Performance of Tools on Test Set

Correct annotation as a fraction of total variants (dependent on identifying the correct transcript)

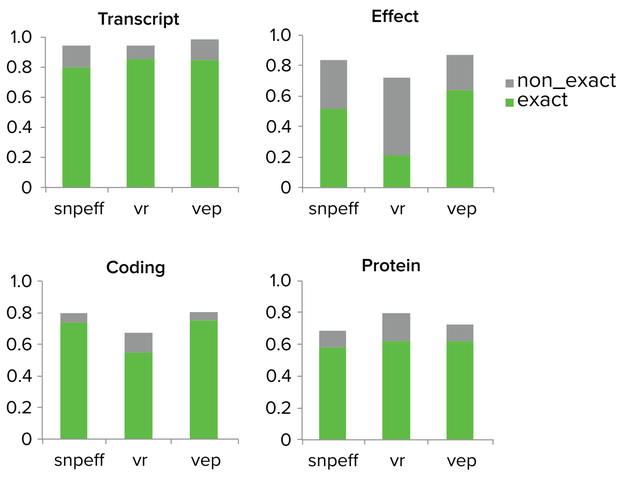


Figure 5b. Performance of Tools on Test Set

Correct annotation as a fraction of variants attempted (independent of transcript annotation set)

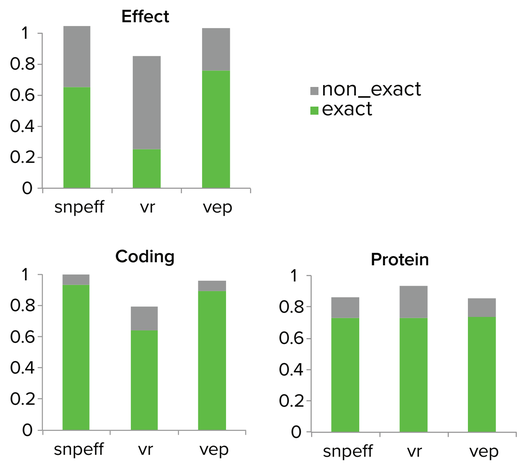


Table 4. Problems (or Bugs) with Variation Reporter

Issue	Input	Output	Reference
Provides conflicting annotations for the same variant	chr9:g.134385436A>G	NM_007171.3:c.752G>A, NM_007171.3:c.752G=	NM_007171.3:c.752G=
When multiple syntax are provided, over-reports effect impact as synonymous	chr7:g.117199644ATCT>A	NM_000492.3:c.1519delAinsATCT, synonymous_codon	NM_000492.3:c.1521_1523delCTT, inframe_deletion
Systematically reports some deletions as insertion/deletions	chr5:g.77396835TTTC>T	NM_003664.4:c.2412delAinsGAAA	NM_003664.4:c.2409_2411delGAA

- Using the Test Set, we identied areas where the tools performed worse than on the ClinVar dataset. For example, Variation Reporter performed poorly on annotating coding syntax and effect impact compared to SnpEff and VEP for deletions and insertions.
- While results for SnpEff and VEP were comparable on the ClinVar dataset, SnpEff slightly outperformed VEP on coding annotation on the Test Set.

Personalis®
Pioneering Genome-Guided Medicine