TransVar: a multilevel variant annotator for precision genomics

To the Editor: To facilitate standardization and reveal inconsistencies in existing variant annotations, we have designed a novel variant annotator, TransVar (http://www.transvar.net), to perform three main functions supporting diverse reference genomes and transcript databases (**Fig. 1a**): (i) forward annotation, which annotates all potential effects of a genomic variant on mRNAs and proteins; (ii) reverse annotation, which traces an mRNA or protein variant to all potential genomic origins; and (iii) equivalence annotation, which, for a given protein variant, searches for alternative protein variants that have an identical genomic origin but are represented on the basis of different isoforms. No robust solutions currently exist for reverse and equivalence annotation, which leads to difficulty in interpreting variants at the protein or mRNA level.

One DNA sequence can code for multiple different mRNAs, and therefore many different proteins. Conversely, a variant identified at the protein or mRNA level may have a non-unique genomic origin. For example, the protein variant *EGFR*:p.L747S, which mediates acquired resistance of non–small cell lung cancer to tyrosine kinase inhibitors[1], can be translated from multiple genomic variants such as chr7:g.55249076_55249077delinsAG and chr7:g.55242470T>C on different isoforms defined on the human reference assembly GRCh37 (**Fig. 1a**). One-to-many, many-to-one and many-to-many relationships among sequence variants at the genomic level and those at transcript and protein levels introduce frequent inconsistencies in current practice when vital information about the annotation process (for example, transcript or isoform I.D.s) is omitted from variant identifiers.

To demonstrate the degree of inconsistency in existing variant data and to evaluate TransVar's forward-annotation functionality, we annotated 964,132 unique single-nucleotide substitutions (SNSs), 3,715 multinucleotide substitutions (MNSs), 11,761 insertions (INSs), 24,595 deletions (DELs) and 166 block substitutions (BLSs) in the Catalogue of Somatic Mutations in Cancer (COSMIC; v67) using TransVar, ANNOVAR[2], VEP[3], snpEff[4] and Oncotator[5] and asked whether the resulting protein identifiers (gene name, protein coordinates and reference amino acid (AA)) matched those in COSMIC. We observed comparable consistency in SNSs and MNSs but variable consistency in INSs, DELs and BLSs from different annotators (**Fig. 1b**, **Supplementary Table 1** and **Supplementary Note 1**). The inconsistencies could be attributed largely to a lack of standardization among variant annotations (codon or AA positions) submitted to COSMIC and among conventions implemented in various annotators. Inconsistency in annotations blurred the lines of evidence for variant-frequency estimation and led to inaccurate determinations of variant function (**Supplementary Note 1**). TransVar revealed hidden inconsistency in these variant annotations by comprehensively identifying alternative annotations in all available transcripts in standard HGVS nomenclature, and thus demonstrated greater consistency in this experiment than achieved with other annotators.

TransVar's reverse annotation can be used to ascertain whether two protein variants have an identical genomic origin, thereby reducing inconsistency in annotation data. It can also show whether a protein variant has non-unique genomic origins and requires caution in genetic and clinical interpretation. We reverse-annotated the protein-level variants in COSMIC and found that even under the constraints imposed by the reference base or AA identity, a sizeable fraction (for example, 11.9% of single-AA substitutions) were associated with multiple genomic variants (**Supplementary Table 2**), if transcripts were not specified. Among 537 variants cited as clinically actionable on the MD Anderson Cancer Center's Personalized Cancer Therapy website (https://pct.mdanderson.org/#/), 78 (14.5%) (for example, *CDKN2A*:p.R87P and *ERBB2*:p.L755_T759del) could be mapped to multiple genomic locations (**Supplementary Table 3**). The reverse-annotation functionality also enabled systematic genomic characterization of variants directly identified from proteomic or RNA-seq data. For example, with just a few minutes of computing time we were able to identify the putative genomic origins of 187,464 (97.69%) protein phosphorylation sites (for example, p.Y308/p.S473 in *AKT1* and p.Y1068/p.Y1172 in *EGFR*) in human proteins[6].

Using the forward- and reverse-annotation features in TransVar can reveal hidden inconsistency and improve the precision of translational and clinical genomics. The tool (methods provided in **Supplementary**

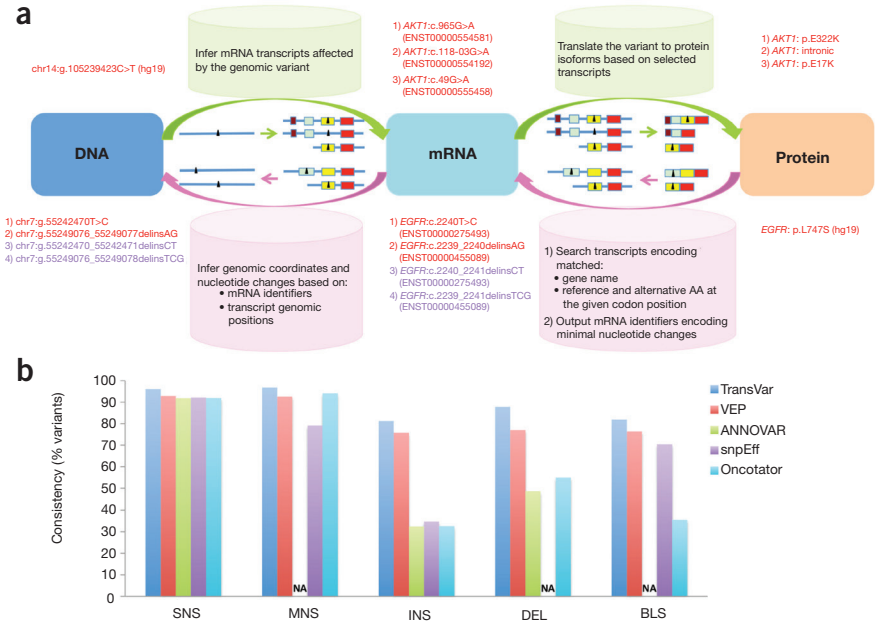

Figure 1 | Overview of TransVar and comparison with other annotation tools. (**a**) TransVar performs forward (green arrows) and reverse (pink arrows) annotation and considers all possible mRNA or protein isoforms available in the user-specified reference genome and transcript databases (colored boxes represent exons). Given a genomic, mRNA or protein variant at one level (black triangles), TransVar can infer associated variants at the other two levels. In reverse annotation, TransVar searches all potential transcripts and reports one variant on each relevant transcript. When there are multiple variants on the same transcript, TransVar reports the variant with minimal nucleotide changes (red text) in addition to alternatives (purple text). (**b**) Consistency of forward annotation among annotation tools. Consistency is represented by the percentage of variants matching protein annotations in COSMIC v67 based on 964,132 unique SNSs, 3,715 MNSs, 11,761 INSs, 24,595 DELs and 166 BLSs. NA, protein-level annotations not available.

Note 2) is available as a user-friendly web interface (http://www.transvar.net) or a downloadable version for batch analysis (**Supplementary Software** and https://bitbucket.org/wanding/transvar).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.3622).

ACKNOWLEDGMENTS

We thank P. Ng and K. Shaw for critical feedback, as well as A. Johnson, A. Bailey, V. Holla, B. Litzenburger, J. Zhang and A. Chang for assistance. This work was supported in part by the US National Institutes of Health (grants CA172652, CA168394, CA083639, CA143883, UL1 TR000371, P50 CA083639, U54 CA112970 and P50 CA098258), the MD Anderson Cancer Center Sheikh Khalifa Ben Zayed Al Nahyan Institute of Personalized Cancer Therapy, the Bosarge Family Foundation, the Mary K. Chapman Foundation, the Michael & Susan Dell Foundation (honoring Lorraine Dell) and the National Cancer Institute Cancer Center (Support Grant P30 CA016672).

AUTHOR CONTRIBUTIONS

K.C. conceived the project. T.C., W.Z. and K.C. designed the studies. W.Z. and T.C. developed the tool and performed the analysis. Z.C. prepared the databases. W.Z., T.C., M.A.R., J.M.M. and C.W. set up the web-application interface. J.Z. and F.M.-B. detected clinical actionable mutations and informed clinical impact. T.C., W.Z., J.N.W., G.B.M. and K.C. interpreted the results and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Wanding Zhou[1,6], Tenghui Chen[1,6], Zechen Chong[1], Mary A Rohrdanz[1], James M Melott[1], Chris Wakefield[1], Jia Zeng[2], John N Weinstein[1,3], Funda Meric-Bernstam[2,4,5], Gordon B Mills[2,3] & Ken Chen[1]

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [2]Khalifa Bin Zayed Al Nahyan Institute of Personalized Cancer Therapy, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [3]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [4]Department of Investigational Cancer Therapy, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [5]Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [6]These authors contributed equally to this work.
e-mail: kchen3@mdanderson.org

1. Yamaguchi, F. *et al. Oncol. Lett.* **7**, 357–360 (2014).
2. Wang, K., Li, M. & Hakonarson, H. *Nucleic Acids Res.* **38**, e164 (2010).
3. McLaren, W. *et al. Bioinformatics* **26**, 2069–2070 (2010).
4. Cingolani, P. *et al. Fly (Austin)* **6**, 80–92 (2012).
5. Ramos, A.H. *et al. Hum. Mutat.* **36**, E2423–E2429 (2015).
6. Hornbeck, P.V. *et al. Nucleic Acids Res.* **40**, D261–D270 (2012).

Avoiding abundance bias in the functional annotation of post-translationally modified proteins

To the Editor: Identification of post-translational modifications (PTMs) by mass spectrometry is biased toward abundant proteins, skewing our understanding of PTMs in favor of readily detected proteins. We developed a method and web tool (https://agotool.sund.ku.dk) to account for this protein-abundance bias in Gene Ontology (GO)-term enrichment analyses of PTM data sets.

GO-term enrichment analysis is frequently used to examine '-omics' data sets for enriched functional terms in a subset of the data set, such as regulated genes or modified proteins[1]. Because the identification of PTMs is biased toward abundant proteins that are more readily detected in the mass spectrometer, GO-enrichment analyses comparing post-translationally modified proteins (referred to here as modified proteins) to unmodified proteins are likely to reveal enriched GO terms associated with abundant proteins, rather than modified proteins specifically.

We mapped thousands of serine and threonine phosphorylation, lysine (N-ε) ubiquitylation, lysine (N-ε) acetylation and lysine (N-ε) succinylation sites in yeast (*Saccharomyces cerevisiae*) and human cervical cancer (HeLa) cells (**Supplementary Tables 1** and **2** and **Supplementary Methods**). Ubiquitylation, acetylation and succinylation were significantly biased toward detection on abundant proteins present in our samples (**Fig. 1a** and **Supplementary Fig. 1a**), although it is also possible that some of the most abundant proteins were not biochemically accessible or present in the cell types analyzed.

We developed a method to account for this abundance bias. GO-term enrichment analysis of PTMs typically involves the use of a statistical test to find significant differences in the frequency of GO terms associated with modified proteins relative to their frequency for the genome or the experimentally observed proteome. To compare modified proteins with an appropriate control group, we applied a protein-abundance-based correction factor to the GO-term associations for the observed proteome. In brief, proteins were binned according to their abundance, and the frequency of GO-term association in each bin was weighted on the basis of the fraction of modified proteins in each bin (**Supplementary Methods**). We developed an open-source, publically accessible web tool, A.GO.TOOL (https://agotool.sund.ku.dk), to perform these analyses.

Using this tool we performed GO-term and UniProt-keyword (KW) enrichment, comparing modified proteins with the genome, observed proteome and abundance-corrected (corrected) proteome (**Fig. 1b** and **Supplementary Data Set 1**). Relative to results for the genome and observed proteome, the number of significantly ($P < 0.01$) enriched GO terms and KWs was decreased when we used the corrected proteome (**Fig. 1b** and **Supplementary Fig. 2**). This reduction was partly attributed to the comparatively small sample size of the corrected proteome (**Supplementary Fig. 3**). Detection of phosphorylation was not abundance-biased in HeLa cells (**Fig. 1a**); therefore, we attributed the reduced association of GO terms and KWs with the corrected proteome (**Fig. 1b**) mostly to the sample size in those analyses (**Supplementary Fig. 3**). The abundance-corrected analysis identified overrepresented (enriched) and underrepresented GO terms and KWs describing specific functions for PTMs, such as "transcription" for acetylation and "proteasome" for ubiquitylation (**Fig. 1c**), as well as organism-specific associations such as prominent enrichment of membrane-linked terms for ubiquitylation in yeast (**Fig. 1c**). Succinylation was not enriched for any GO terms or KWs in yeast, regardless of the sample size (**Fig. 1b** and **Supplementary Fig. 3**). This striking result is consistent with untargeted, nonenzymatic succinylation[2,3]. An independent analysis of bacterial acetylation similarly showed no significantly enriched GO terms[4], consistent with prominent nonenzymatic acetylation in bacteria[5]. Acetylation can occur nonenzymatically in eukaryotes[2,6,7], and it is also catalyzed by acetyltransferases that primarily regulate gene transcription. After correcting for abundance bias, we found that GO terms and KWs describing transcription and related processes were significantly enriched for acetylated proteins in yeast and HeLa cells (**Fig. 1d,e** and **Supplementary Data Set 1**). In contrast, KWs associated with central metabolism were not significantly enriched in analyses using the corrected proteome (**Fig. 1f**), and the decreased significance was not attributed to sample size (**Supplementary Fig. 4**).

Previously published GO-enrichment analyses, including several of our own and many not cited here, show that acetylation and