Published in final edited form as:

Hum Mutat. 2015 April ; 36(4): E2423–E2429. doi:10.1002/humu.22771.

Oncotator: cancer variant annotation tool

Alex H. Ramos[1,2,3,4,#], **Lee Lichtenstein**[1,#], **Manaswi Gupta**[1], **Michael S. Lawrence**[1], **Trevor J. Pugh**[1,2,3], **Gordon Saksena**[1], **Matthew Meyerson**[1,2,3,*], **Gad Getz**[1,3,4,*]

[1]Cancer Program, Broad Institute, Cambridge, MA 02142, USA

[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA

[3]Department of Pathology, Harvard Medical School, Boston, MA 02115, USA

[4]Cancer Center and Department of Pathology, Massachusetts General Hospital Boston, MA 02114, USA

Abstract

Oncotator is a tool for annotating genomic point mutations and short nucleotide insertions/deletions (indels) with variant- and gene-centric information relevant to cancer researchers. This information is drawn from 14 different publicly available resources that have been pooled and indexed, and we provide an extensible framework to add additional data sources. Annotations linked to variants range from basic information, such as gene names and functional classification (e.g. missense), to cancer-specific data from resources such as the Catalogue of Somatic Mutations in Cancer (COSMIC), the Cancer Gene Census, and The Cancer Genome Atlas (TCGA). For local use, Oncotator is freely available as a python module hosted on Github (https://github.com/broadinstitute/oncotator). Furthermore, Oncotator is also available as a web service and web application at http://www.broadinstitute.org/oncotator/.

INTRODUCTION

Variant annotation, the aggregation and reporting of data relevant to a given genomic alteration, is a key step in a sequencing data analysis pipeline and is crucial for subsequent interpretation of detected variants. Genome sequencing of cancer samples typically reveal thousands to tens of thousands of somatic mutations per tumor that are often unique to the individual tumor (equivalent to ‘singletons’ in a germline analysis) (Lawrence et al., 2013). Therefore, researchers rely on annotations to filter variants to a subset of alterations that are most important to a given study or application. At the most basic level, variant annotations help researchers identify the genes, transcripts, and genomic regions pertaining to a given variant, as well as predict the impact an alteration has on the translated protein product of a gene. With the emergence of large databases of germline (Sherry et al., 2001; Landrum et al., 2013; NHLBI GO Exome Sequencing Project, 2014) and somatic (Forbes et al., 2010) variation, synthesis of all available clinical and biological information for single variants

Contact: gadgetz@broadinstitute.org.
[#]These authors contributed equally.
[*]To whom correspondence should be addressed.

greatly empowers researchers to distinguish driver mutations from passengers (Imielinski et al., 2014), link specific variants to patient phenotypes (Van Allen et al., 2013; Biesecker and Green, 2014), and uncover unexpected oncogenic mechanisms shared across diseases (Lawrence et al., 2013). For example, variant annotations such as the frequency of a mutation in published cancer genomic studies, or whether a significant functional effect is predicted by algorithms such as Polyphen-2 or SIFT (Kumar et al., 2009; Adzhubei et al., 2010), can be utilized to interpret and prioritize variants. For clinicians, variant annotations can be of immense aid for clinical interpretation of variants as annotations can identify genetic events associated with cancer prognosis/diagnosis or drug sensitivity/resistance (Van Allen et al., 2013; Biesecker and Green et al., 2014).

Numerous tools, such as ANNOVAR, SnpEff, and Variant Effect Predictor, exist for annotating sequencing variants; however, many were developed for general non-cancer applications (Le Pera et al., 2010; McLaren et al., 2010; Wang et al., 2010; Cingolani et al., 2011; Sana et al., 2011). Although cancer sequencing studies use many of these tools, variants may lack cancer-specific annotations that can aid in downstream interpretation.

Here we report Oncotator, a cancer variant annotation pipeline implemented as a command line tool, as well as a web application, which provides both an interactive user interface and a programmatic web service. As currently deployed, Oncotator allows users to annotate variants with a pre-packaged bundle of cancer-relevant information in a single step. Oncotator has been used internally in the Broad Institute's Cancer Genome Analysis pipeline since 2011, resulting in its use in over 20 published cancer studies, including several large scale (>100 tumors) efforts conducted by TCGA, NHGRI, TARGET, and the Slim Initiative for Genomic Medicine collaboration (Bass et al., 2011; Berger et al., 2011; Cancer Genome Atlas Research Network, 2011, 2013, 2014; Chapman et al., 2011; Hammerman et al., 2011; Stransky et al.. 2011; Wang et al., 2011; Banerji et al., 2012; Barbieri et al., 2012; Barretina et al., 2012; Berger et al., 2012; Hodis et al., 2012; Imielinski et al., 2012; Lee et al., 2012; Lohr et al., 2012; Pugh et al., 2012, 2013, 2014; Ciriello et al., 2013; Francis et al., 2013; Lawrence et al., 2013; Ojesina et al., 2013). The goal of this article is to make the scientific community aware of the first public release of Oncotator (version 1.3) that is free to non-profit users. In the past two years, large parts of Oncotator were refactored to support: (i) highly optimized annotations; (ii) customizable data sources; and (iii) deployment outside of the Broad Institute environment.

METHODS

As a starting point for annotation, Oncotator requires the genomic position, reference allele, and variant allele as input in TSV, VCF (Danecek et al., 2011), or muTect call_stats (Cibulskis et al., 2013) formats. Variants are currently annotated with data from 14 different resources (Table 1), described briefly below. Annotated variants can be output in TCGA MAF or VCF formats, regardless of which input format is used (with the unintended consequence that researchers often use the tool for format conversion).

Oncotator uses a local indexed database of reference transcripts derived from GENCODE to map variants to specific genes (Harrow et al., 2012). Each variant is assigned a "variant

classification" (e.g. "Splice_Site" or "Nonsense_Mutation") based on the mutation's position relative to an overlapping gene and the expected consequence, if any, that the mutation has on a translated protein product. The model of reference transcript selection used is user-defined by a command line argument or can be left to automatically use the model with the greatest deleterious effect. Variant classification terms defined by the TCGA are used and nomenclature adheres to specifications defined by the Human Genome Variation Society (http://www.hgvs.org/mutnomen) (Dunnen et al., 2000).

In addition to basic transcript annotations described above, Oncotator will annotate variants with annotations derived from sources that can be beneficial to researchers looking to prioritize variants. To identify common Single Nucleotide Polymorphism (SNP) variants (which are less likely to contribute to tumorigenesis), Oncotator utilizes data from dbSNP, 1000 Genomes Project, and National Heart, Lung, and Blood Institute's Exome Sequence Project (Sherry et al., 2001; 1000 Genomes Project Consortium. 2010; NHLBI GO Exome Sequencing Project, 2014). Oncotator can also annotate variants with the local GC content (within 100 base-pairs, by default) and surrounding nucleotide context (within 10 base-pairs, by default). Such annotations can be helpful for identifying biological mutational processes with sequence-specific mutation (Lawrence et al., 2013; Alexandrov et al., 2013) or artifactual mutation biases such as oxidation of guanine bases during sequencing library construction (OxoG) (Costello et al., 2013).

Predicting the functional impact of somatic mutations in cancer can be aided by mapping coding DNA sequence variants in genes onto amino acid sequences and proteins they encode. For example, knowledge of the specific protein regions that variants affect can be used to identify particular protein domains or active sites that are enriched for mutations across multiple samples or even across genes containing similar domains. To this end, Oncotator can annotate genomic variants with protein-specific annotations derived from UniProt human protein sequence records (UniProt Consortium, 2011). Oncotator maps genomic variants to protein position-based annotations derived from the feature table section of a UniProt record. Protein annotations added include "region" (e.g. protein kinase domain), "site" (e.g. ATP binding site), "natural variation" (e.g. Y → F in Pfeiffer syndrome), and "experimental" (e.g. Y → F: 50% decrease in interaction with PIK3C2B) data, if available. Furthermore, UniProt records are utilized to derive Gene Ontology (GO) annotations, describing the biological process, cellular component, and molecular function of a gene; and DrugBank annotations, pertaining to small molecules known to target the protein of interest (Knox et al., 2011). Through the dbNSFP (Liu et al., 2011) Oncotator datasource, variants can also be annotated with pre-computed results derived from many functional prediction and conservation score algorithms (PolyPhen-2, SIFT, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, GERP++ and PhyloP), which can be used to classify variants most likely to have an impact on a protein's function (Siepel et al., 2006; Chun and Fay, 2009; Garber et al., 2009; Kumar et al., 2009; Adzhubei et al., 2010; Davydov et al., 2010; Schwarz et al., 2010; Reva et al., 2011; Shihab et al., 2013).

Oncotator also annotates variants with data from several cancer-specific resources that may aid in interpreting variants. Using data from COSMIC, Oncotator identifies variants reported in published studies and reports their observed frequency across all cancers and within each

tissue type (Forbes et al., 2010). Overlapping breakpoint and fusion genes in COSMIC are also provided. Cancer researchers can also benefit from knowledge of relevant cancer cell line models in which to perform follow-up in vitro experiments with. To this end, Oncotator utilizes data from the Cancer Cell Line Encyclopedia to identify if a variant has been previously observed in a cell line (Barretina et al., 2012) (http://www.broadinstitute.org/ccle/home). Other cancer-specific resources utilized include the Cancer Gene Census (Futreal et al., 2004), ClinVar (Landrum et al., 2014), the Familial Cancer Database (http://www.familialcancerdatabase.nl/), and a curated set of DNA repair genes (Wood et al., 2005).

IMPLEMENTATION

Oncotator is available as a command line tool written in the Python programming language (https://www.python.org/). This tool is recommended for advanced users and is ready for inclusion into automated pipelines, since the annotation options, selection of data sources, and file formats are more flexible. The Oncotator software is an annotation framework that is broken into a three-stage workflow: (i) convert the input data into an internal model of mutations; (ii) annotate the mutation objects with a collection of pre-processed datasources (which can be locus-, variant- or gene-specific); and (iii) render the mutations to the specified output format (VCF or MAF). The software architecture encapsulates each step, which allows easy implementation of input and output formats by decoupling file formats from the actual annotation engine. The encapsulation also eased the development of hundreds of automated test modules, some testing hundreds of scenarios, that allow developers to make code changes and be confident that their changes have not unintentionally broken previous functionality (regression tests).

We recognize that researchers would like to extend variant annotations beyond the current available datasources in Oncotator. Therefore, we included in Oncotator tools for creating new datasources from TSV, VCF, and GTF files. Most of the default Oncotator datasources were created using these tools. Users are encouraged to contribute to the project via a publicly available Github repository (https://github.com/broadinstitute/oncotator). Although the tool was initially developed for cancer researches, Oncotator can address non-cancer needs and additional non-cancer datasources can be easily introduced. Periodically, we make updated and new datasources available, as part of the versioned default corpus. In the future, we plan to add datasources specific to whole genome sequencing analysis, such as conservation scores outside of the exome, as well as add functionality for annotation of genomic regions.

Oncotator is also available as a web application at http://www.broadinstitute.org/oncotator/. Users can input a tab-delimited text file containing genomic coordinates and allele genotypes for each variant. Annotation results are presented as interactive tables. Users can also download a tab-delimited file containing multiple columns corresponding to the different annotations that are aggregated. The Oncotator web service is implemented using a REST-like architecture to facilitate integration with existing applications and pipelines. Users can also retrieve variant annotations programmatically using HTTP requests in the form http://www.broadinstitute.org/oncotator/mutation/

<chromosome>_<start_position>_<end_position>_<reference_allele>_<observed_allele>. Results are returned as JSON objects (http://json.org/) which can be easily parsed by users.

ACKNOWLEDGEMENTS

We would like to thank our colleagues from the Broad Institute Cancer Program, The Cancer Genome Atlas project, and beta test users who supported the development of Oncotator and provided valuable feedback.

REFERENCES

1000 Genomes Project Consortium. 2010 A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–73. [PubMed: 20981092]

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010 A method and server for predicting damaging missense mutations. Nat Methods 7(4):248–249. [PubMed: 20354512]

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL, Boyault S, Burkhardt B, et al. 2013 Signatures of mutational processes in human cancer. Nature 500(7463):415–421. [PubMed: 23945592]

Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L, Cortes ML, Fernandez-Lopez JC, et al. 2012 Sequence analysis of mutations and translocations across breast cancer subtypes. Nature 486(7403):405–409. [PubMed: 22722202]

Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, White TA, Stojanov P, Van Allen E, Stransky N, Nickerson E, Chae SS, et al. 2012 Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet 44(6):685–689. [PubMed: 22610119]

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, Reddy A, Liu M, et al. 2012 The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483(7391):603–607. [PubMed: 22460905]

Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, Sougnez C, Voet D, Saksena G, Sivachenko A, Jing R, Parkin M, et al. 2011 Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat Genet 43(10):964–968. [PubMed: 21892161]

Berger MF, Hodis E, Heffernan TP, Deribe YL, Lawrence MS, Protopopov A, Ivanova E, Watson IR, Nickerson E, Ghosh P, Zhang H, Zeid R, et al. 2012 Melanoma genome sequencing reveals frequent PREX2 mutations. Nature 485(7399):502–506. [PubMed: 22622578]

Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, Onofrio R, Carter SL, et al. 2011 The genomic complexity of primary human prostate cancer. Nature 470(7333):214–220. [PubMed: 21307934]

Biesecker LG, Green RC. 2014 Diagnostic Clinical Genome and Exome Sequencing. N Engl J Med 370(25):2418–2425. [PubMed: 24941179]

Cancer Genome Atlas Research Network. 2011 Integrated genomic analyses of ovarian carcinoma. Nature 474(7353):609–615. [PubMed: 21720365]

Cancer Genome Atlas Network. 2013 Comprehensive molecular characterization of human colon and rectal cancer. Nature 487(7407):330–337.

Cancer Genome Atlas Network. 2014 Comprehensive molecular characterization of urothelial bladder carcinoma. Nature 507(7492):315–22. [PubMed: 24476821]

Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M, Anderson KC, Ardlie KG, et al. 2011 Initial genome sequencing and analysis of multiple myeloma. Nature 471(7339):467–472. [PubMed: 21430775]

Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, Nickerson E, Auclair D, Li L, Place C, Dicara D, Ramos AH, et al. 2012 A Landscape of Driver Mutations in Melanoma. Cell 150(2):251–263. [PubMed: 22817889]

Chun S, Fay JC. 2009 Identification of deleterious mutations within three human genomes. Genome Res 19(9):1553–1561. [PubMed: 19602639]

Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. 2013 Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31(3):213–219. [PubMed: 23396013]

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012 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 (Austin) 6(2):80–92. [PubMed: 22728672]

Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. 2013 Emerging landscape of oncogenic signatures across human cancers. Nat Genet 45(10):1127–1133. [PubMed: 24071851]

Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, Meldrim JC, Fostel JL, Friedrich DC, Perrin D, Dionne D, Kim S, Gabriel SB, et al. 2013 Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res 41(6):e67. [PubMed: 23303777]

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. 2011 The variant call format and VCFtools. Bioinformatics 27(15):2156–2158. [PubMed: 21653522]

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010 Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6(12):e1001025. [PubMed: 21152010]

den Dunnen JT, Antonarakis SE. 2000 Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 15(1):7–12. [PubMed: 10612815]

Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, et al. 2010 COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 39(Database issue):D945–50. [PubMed: 20952405]

Francis JM, Kiezun A, Ramos AH, Serra S, Pedamallu CS, Qian ZR, Banck MS, Kanwar R, Kulkarni AA, Karpathakis A, Manzo V, Contractor T, et al. 2013 Somatic mutation of CDKN1B in small intestine neuroendocrine tumors. Nat Genet 45(12):1483–1486. [PubMed: 24185511]

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004 A census of human cancer genes. Nat Rev Cancer 4(3):177–183. [PubMed: 14993899]

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009 Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 25(12):i54–62. [PubMed: 19478016]

Hammerman PS, Sos ML, Ramos AH, Xu C, Dutt A, Zhou W, Brace LE, Woods BA, Lin W, Zhang J, Deng X, Lim SM, et al. 2011 Mutations in the DDR2 Kinase Gene Identify a Novel Therapeutic Target in Squamous Cell Lung Cancer. Cancer Discov 1(1):78–89. [PubMed: 22328973]

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, et al. 2012 GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22(9):1760–1774. [PubMed: 22955987]

Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, Cho J, Suh J, Capelletti M, Sivachenko A, Sougnez C, Auclair D, et al. 2012 Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell 150(6):1107–1120. [PubMed: 22980975]

Imielinski M, Greulich H, Kaplan B, Araujo L, Amann J, Horn L, Schiller J, Villalona-Calero MA, Meyerson M, Carbone DP. 2014 Oncogenic and sorafenib-sensitive ARAF mutations in lung adenocarcinoma. J Clin Invest 124(4):1582–1586. [PubMed: 24569458]

Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, et al. 2011 DrugBank 3.0: a comprehensive resource for ‘omics’ research on drugs. Nucleic Acids Res 39(Database issue):D1035–41. [PubMed: 21059682]

Kumar P, Henikoff S, Ng PC. 2009 Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4(8):1073–1081. [PubMed: 19561590]

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, et al. 2013 Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499(7457):214–8. [PubMed: 23770567]

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014 ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42(Database issue):D980–985. [PubMed: 24234437]

Le Pera L, Marcatili P, Tramontano A. 2010, PICMI: mapping point mutations on genomes. Bioinformatics 26(22):2904–2905. [PubMed: 20940168]

Lee RS, Stewart C, Carter SL, Ambrogio L, Cibulskis K, Sougnez C, Lawrence MS, Auclair D, Mora J, Golub TR, Biegel JA, Getz G, Roberts CW. 2012 A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. J Clin Invest 122(8):2983–2988. [PubMed: 22797305]

Liu X, Jian X, Boerwinkle E. 2013 dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat 34(9):E2393–402. [PubMed: 23843252]

Lohr JG, Stojanov P, Lawrence MS, Auclair D, Chapuy B, Sougnez C, Cruz-Gordillo P, Knoechel B, Asmann YW, Slager SL, Novak AJ, Dogan A, et al. 2012 Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proc Natl Acad Sci U S A 109(10):3879–3884. [PubMed: 22343534]

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010 Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26(16):2069–2070. [PubMed: 20562413]

NHLBI GO Exome Sequencing Project. 7, 2014 Exome Variant Server (URL: http://evs.gs.washington.edu/EVS/).

Ojesina AI, Lichtenstein L, Freeman SS, Pedamallu CS, Imaz-Rosshandler I, Pugh TJ, Cherniack AD, Ambrogio L, Cibulskis K, Bertelsen B, Romero-Cordoba S, Treviño V, et al. 2013 Landscape of genomic alterations in cervical carcinomas. Nature 506(7488):371–5. [PubMed: 24390348]

Pugh TJ, Morozova O, Attiyeh EF, Asgharzadeh S, Wei JS, Auclair D, Carter SL, Cibulskis K, Hanna M, Kiezun A, Kim J, Lawrence MS, et al. 2013 The genetic landscape of high-risk neuroblastoma. Nat Genet 45(3):279–284. [PubMed: 23334666]

Pugh TJ, Weeraratne SD, Archer TC, Pomeranz Krummel DA, Auclair D, Bochicchio J, Carneiro MO, Carter SL, Cibulskis K, Erlich RL, Greulich H, Lawrence MS, et al. 2012 Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. Nature 488(7409):106–110. [PubMed: 22820256]

Pugh TJ, Yu W, Yang J, Field AL, Ambrogio L, Carter SL, Cibulskis K, Giannikopoulos P, Kiezun A, Kim J, McKenna A, Nickerson E, et al. 2014 Exome sequencing of pleuropulmonary blastoma reveals frequent biallelic loss of TP53 and two hits in DICER1 resulting in retention of 5p-derived miRNA hairpin loop sequences. Oncogene 0:[Epub ahead of print].

Reva B, Antipin Y, Sander C. 2011 Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 39(17):e118. [PubMed: 21727090]

Sana ME, Iascone M, Marchetti D, Palatini J, Galasso M, Volinia S. 2011 GAMES identifies and annotates mutations in next-generation sequencing projects. Bioinformatics 27(1):9–13. [PubMed: 20971986]

Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. 2010 MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7(8):575–576. [PubMed: 20676075]

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001 dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308–11. [PubMed: 11125122]

Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR. 2013 Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Human Mutat 34(1):57–65.

Siepel A, Pollard KS, Haussler D. 2006 New methods for detecting lineage-specific selection. Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006):190–205.

Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, Kryukov GV, Lawrence MS, Sougnez C, McKenna A, Shefler E, Ramos AH, et al. 2011 The mutational landscape of head and neck squamous cell carcinoma. Science 333(6046):1157–1160. [PubMed: 21798893]

Consortium UniProt. 2011 Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res 39(Database issue):D214–9. [PubMed: 21051339]

Van Allen EM, Wagle N, Levy MA. 2013 Clinical Analysis and Interpretation of Cancer Genome Data. J Clin Oncol 31(15):1825–1833. [PubMed: 23589549]

Wang K, Li M, Hakonarson H. 2010 ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16):e164. [PubMed: 20601685]

Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, Werner L, Sivachenko A, DeLuca DS, Zhang L, Zhang W, Vartanov AR, et al. 2011 SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. N Engl J Med 365(26):2497–2506. [PubMed: 22150006]

Wood RD, Mitchell M, Lindahl T. 2005, Human DNA repair genes, 2005. Mutat Res 577(1–2):275–83. [PubMed: 15922366]

Table 1.

Oncotator datasources

Annotation Category	Resource	URL	Comments
Genomic	GENCODE	http://www.gencodegenes.org/	GENCODE/ENSEMBL transcripts and annotations for hg19
	ref_context		Can be used for artifact inference
	gc_content		Can be used for artifact inference
	Human DNA Repair Genes	http://sciencepark.mdanderson.org/labs/wood/DNA_Repair_Genes.html	Alteration in such genes can help explain higher overall mutation rates in specific samples
Protein	UniProt	http://www.uniprot.org/	Includes Drugbank & GO annotations
	dbNSFP	https://sites.google.com/site/jpopgen/dbNSFP	Contains pre-computed conservation scores, prediction classifications, and other information
Cancer Variant	COSMIC	http://www.sanger.ac.uk/genetics/CGP/cosmic/	
	Cancer Gene Census	http://www.sanger.ac.uk/genetics/CGP/Census/	
	CCLE	http://www.broadinstitute.org/ccle/home	Cancer cell line annotations. Can be used to identify cell line models containg variants of interest
	Familial Cancer Database	http://www.familialcancerdatabase.nl/	
	ClinVar	http://www.ncbi.nlm.nih.gov/clinvar/	
Non-Cancer Variant	dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/	1000G variants are denoted
	1000 Genomes	http://www.1000genomes.org/data	
	NHLBI GO Exome Sequencing Project (ESP)	https://esp.gs.washington.edu/drupal/	

Hum Mutat. Author manuscript; available in PMC 2020 July 10.