Seshat: A Web service for accurate annotation, validation, and analysis of TP53 variants generated by conventional and next-generation sequencing

Tuomas Tikkanen, Bernard Leroy, Jean Louis Fournier, Rosa Ana Risques, Jitka Malcikova, Thierry Soussi

Tuomas Tikkanen[1], Bernard Leroy[2], Jean Louis Fournier[2], Rosa Ana Risques[3], Jitka Malcikova[4,5] and Thierry Soussi[2,6,7,8]

Seshat: a web service for accurate annotation, validation and analysis of *TP53* variants generated by conventional and next-generation sequencing.

[1] Genevia Technologies, Hämeenkatu 14 C 33, 33100 Tampere, Finland

[2] Sorbonne Université, UPMC Univ Paris 06, F- 75005 Paris, France

[3] Department of Pathology, University of Washington, Seattle, WA, USA

[4] Department of Internal Medicine – Hematology and Oncology, University Hospital Brno and Medical Faculty, Masaryk University, Brno, Czech Republic

[5] Central European Institute of Technology, Masaryk University, Brno, Czech Republic

[6] Department of Oncology-Pathology, Cancer Center Karolinska (CCK), Karolinska Institutet, Stockholm, Sweden

[7] INSERM, U1138, Centre de Recherche des Cordeliers, Paris, France

[8]: Correspondance: thierry.soussi@ki.se

Keywords: TP53 variants; HGVS variant nomenclature, variant annotation, database

Grant Sponsor: Radiumhemmets Forskningsfonder and the Swedish Cancer Society (Cancerfonden) to TS.

Abstract

Accurate annotation of genomic variants in human diseases is essential to allow personalized medicine. Assessment of somatic and germline *TP53* alterations has now reached the clinic and is required in several circumstances such as the identification of the most effective cancer therapy for patients with chronic lymphocytic leukemia (CLL). Here we present Seshat, a web service for annotating *TP53* information derived from sequencing data. A flexible framework allows the use of standard file formats such as Mutation Annotation Format (MAF) or Variant Call Format (VCF), as well as common TXT files. Seshat performs accurate variant annotations using the the Human Genome Variation Society (HGVS) nomenclature and the stable *TP53* genomic reference provided by the Locus Reference Genomic (LRG). In addition, using the 2017 release of the UMD_TP53 database, Seshat provides multiple statistical information for each *TP53* variant including database frequency, functional activity or pathogenicity. The information is delivered in standardized output tables that minimize errors and facilitate comparison of mutational data across studies. Seshat is a beneficial tool to interpret the ever-growing *TP53* sequencing data generated by multiple sequencing platforms and it is freely available via the TP53 website, http://p53.fr or directly at http://vps338341.ovh.net/.

Introduction

The clinical utility of *TP53* alteration analysis has now been clearly established in several circumstances (Leroy et al., 2017). Firstly, somatic *TP53* status is used in routine clinical practice in several types of cancer such as chronic lymphocytic leukemia (CLL), acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS), in order to identify patients likely to benefit from specific treatment (Döhner et al., 2017; Pospisilova et al., 2012). Secondly, it has been clearly established that germline *TP53* variants are frequent in familial cancer syndromes, such as Li-Fraumeni syndrome (LFS) or in families with hereditary breast and ovarian cancer, and surveillance of individuals with an identified germline *TP53* mutation is highly beneficial to improve the likelihood of early tumor detection and subsequently improved outcomes (Ballinger, Mitchell, & Thomas, 2015). Thirdly, more than 400 clinical trials requiring stratification based on *TP53* status are currently underway (Khoo, Hoe, Verma, & Lane, 2014). Fourthly, multiple drugs targeting mutant TP53 have been developed and are entering clinical trials (Bykov & Wiman, 2014). Fifthly, current efforts directed at cancer detection using liquid biopsies are revealing the presence of low frequency, aging-related, *TP53* variants in individuals without cancer, highlighting the need to understand the nature of these variants in order to develop better strategies for cancer detection (Krimmel et al., 2016; Fernandez-Cuesta et al., 2016).

Finally, it should be stressed that the diagnosis of *TP53* alteration will also become mandatory in stem cell research, as a recent study has shown that a significant number of human embryonic stem cell (hES cell) lines, including lines prepared for potential clinical use, contain somatic *TP53* variants that have arisen during the culturing process (Merkle et al., 2017; Trounson, 2017).The assessment of *TP53* alterations must, therefore satisfy the quality requirements for clinical diagnostic tests used in personalized medicine (Leroy et al., 2017).

In the era of next-generation sequencing (NGS), *TP53* mutation analysis (as well as analysis of other genes) raises a number of issues that were not present in the age of Sanger sequencing. The first problem is related to the huge amount of information processed by multiple independent pipelines, which prevents extensive manual curation. Bioinformatics pipelines used for the analysis of NGS data comprise numerous steps, such as de-multiplexing, read alignment, de-duplication, base calibration, variant calling, filtering, and annotation. There is no single “gold-standard” algorithm at the present time and laboratories often use multiple in-house and/or commercial software for analysis, each devoted to a specific step in the pipeline (Goodwin, McPherson, & McCombie, 2016). Furthermore, all genes are treated simultaneously by the same pipeline, preventing the fine-tuning that could be performed when small numbers of genes and/or patients were analysed by conventional Sanger sequencing (**Figure 1)**.

A second problem is associated with the dramatic increase of transcriptome complexity and the discovery that up to 95% of human genes undergo splicing in a developmental, tissue-specific or signal transduction-dependent manner (Chen & Manley, 2009). The most challenging aspect of variant annotation is the conversion of genomic coordinates (i.e., chromosome and position) to the corresponding cDNA and/or amino acid coordinates. This issue is the source of major problems at the variant annotation step despite the fact that it is critical for an accurate description of the variant and its translation into a clinical decision. In a recent survey of several cancer mutation databases, Yen et al. revealed important inconsistencies in variant representation across annotation tools and databases, an observation that we have reported many times for the *TP53* gene (Yen et al., 2017; Soussi, Leroy, & Taschner, 2014). To solve this general confounding situation, The Locus Reference Genomic (LRG) consortium including European Bioinformatics Institute (EBI), the National Center for Biotechnology Information (NCBI) and Human Genome Variation Society (HGVS) as well as LSDB (Locus Specific Database) curators has designed a

reference system that would allow consistent and unambiguous reporting of variants in clinically relevant loci. LRG provides stable reference sequences and a coordinate system for permanent and unambiguous reporting of disease-causing variants [Dalgleish et al., 2010; MacArthur et al., 2014]. LRGs already cover 1027 genes associated with noncancerous or cancerous diseases (http://www.lrg-sequence.org/).

This transcriptomic complexity has not spared the *TP53* gene and multiple translated and untranslated RNAs have been identified (Khoury & Bourdon, 2010). Unfortunately, annotations of all these transcripts is highly heterogeneous among the various databases. The NCBI website describes a *TP53* gene with 12 exons transcribed in 15 different RNAs (annotation release 108). On the other hand, Ensembl reports 28 different transcripts transcribed from 12 exons of *TP53* gene, but the exon nomenclature between NCBI and Ensembl is not similar. In order to circumvent this problem, a joint effort from *TP53* specialists resulted in the release of a stable *TP53* reference sequence, LRG 321, containing the genomic sequence from human genome build GRCh37.p13 (ftp://ftp.ebi.ac.uk/pub/databases/lrgex/LRG 321.xml). These annotations with precise labels and coordinates of eight different *TP53* transcripts (t1–t8) and 12 protein isoforms (p1 and p3 to p13) should be preferred to the RefSeq identifier pairs provided by the NCBI for genome build GRCh37.p13 or the various transcripts described by Ensembl. The choice of these transcripts and proteins was based on our current knowledge on *TP53* expression issued from various experimental data and validated by the consortium of *TP53* specialists (Joruiz & Bourdon, 2016). We have found numerous studies describing *TP53* variants using either cDNA or protein nomenclature with unidentified mix-ups of references. For example, description of *TP53* variants LRG_321p1:p.R175H as either p.R175H or p.R136H in the same table prevents the user from determining the true reference of the variant and deducing the location of the mutation in the genome. Using

the same reference for the description of all variants in a single output should be mandatory and will at least partially circumvent this problem.

Variant annotation must be performed according to nomenclature guidelines from Human Genome Variation Society (HGVS, available at www.hgvs.org) (Taschner & den Dunnen, 2011). Details regarding the reference of the gene as well as the mRNA and protein are mandatory and should always be included in the variant description. Although this description is relatively easy for SNV, it is more difficult for insertions, deletions, duplications or more complex events such as an insertion with a deletion. The HGVS system recommends right-aligned (shifting the start position of the variant to the 3´ end until it is no longer possible to do so) representation of sequence variants, but the lack of standardization of left/right alignment significantly affects variant localization leading to incorrect variant nomenclature (Taschner & den Dunnen, 2011).

We have also observed that nucleotide duplications are often considered to be insertions. In the *TP53* gene, more than 70% of insertions are indeed nucleotides duplications localized in polynucleotide tracts pointing to a defect in postreplicative DNA repair.

Nomenclature inconsistencies are also observed for tandem mutations, i.e., CC>TT double substitutions, typical for mutagenesis associated with pyrimidine dimers caused by UV exposure and frequently observed in skin cancer. Although they should be described as single mutagenesis events they are often considered as two independent juxtaposing single nucleotide variants, each one leading to a different amino acid substitution. When this event occurs across two juxtaposing codons, the modification of the first codon at the third position often leads to a synonymous modification which is removed by many analytical pipelines as synonymous variants are not considered as pathogenic (Soussi, Taschner, & Samuels, 2017).

Variant annotation is critical for further analysis, as it constitutes the link between identified variants and the multiple analytical databases used to infer relevant information associated with the variant **(Figure 1).** Inaccurate annotations can lead to erroneous interpretation with unknown consequences.

Created in 1989 and updated regularly, the UMD_TP53 database is the most updated repository of *TP53* variants (https://p53.fr/). The current version (Oct 2017), consists of 80,406 samples including recent tumor sequencing data from The International Cancer Genome Consortium (ICGC, http://icgc.org/), The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/) and the Memorial Sloan-Kettering Cancer Center (MSKCC, https://www.synapse.org/#!Synapse:syn7222066/wiki/405659). The UMD_TP53 database is the only *TP53* mutation database that has been curated for sequencing artefacts (Edlund et al., 2012).

We introduce Seshat, a novel portal dedicated to the specific annotation of *TP53* variants based on the UMD_TP53 database (https://p53.fr/tp53-database/seshat). The various features of the portal are as follows: (1) importation of variant data from VCF or MAF files generated by NGS, as well as CSV files with user data, (2) annotation of variants, compilation of information from relevant databases, and statistical analyses based on the UMD_*TP53* database; and (3) generation of output tables with accurate nomenclature, as well as *TP53*-specific information such as pathogenicity or frequency in various types of cancer.

Material and methods

The 2017 release of the UMD_*TP53* database

The current version of Seshat is based on the 2017 release of the UMD_TP53 database that includes 80,406 alterations identified in tumors, cell lines (somatic mutations) or in patients with hereditary cancer (germline mutations) (database freeze Oct 2017). These alterations can be grouped into 6,874 different *TP53* variants (**Figure 2)** (Leroy et al., 2017).

Missense variants are the most common variant type, encompassing 73% of mutations found in human tumors. Some of these variants are highly prevalent such as NP_000537.3:p.R175H (4.2%), NP_000537.3:p.R1248Q (3.2%), or NP_000537.3:p.R273H (2.9%), which are hotspot mutations found at CpG sites. In mammalian cells, the cytosine in this dinucleotide is very often methylated and it has been shown that the 42 CpG sites of the *TP53* gene are methylated in normal tissue. Deamination of 5-methylcytosine leads to mutation much more often than does deamination of cytosine leading to mutational hotspots in mammalian genomes. Missense variants represent only 31% of the unique *TP53* variants due to the fact that frameshift variants are highly diverse (57% of all variants). Frameshift variants, however, are not abundant in human tumors and correspond to only 12% of the events reported in this large database (**Figure 2)**.

The database has been fully analyzed using Mutalyzer to be compliant with HGVS recommendations (Wildeman, van Ophuizen, den Dunnen, & Taschner, 2008). Multiple deletion variants in homopolymeric tracts have been combined as a single variant using the 3’ rule. We have also observed that more than 70% of insertions are actually nucleotide duplications.

For each variant, UMD_TP53 includes its frequency in the database and the residual activity of the mutant variant based on the transcriptional activity assay for 3,000 variants performed by Kato et al. (Kato et al., 2003). For an easier interpretation, functional data are given as a percentage compared to wild-type protein. As shown in **Supp. Figure S1**, database frequency and loss of the transcriptional activity are highly correlated and are the most potent pathogenicity predictors.

Each variant of the UMD_TP53 now includes multiple annotations derived from databases such as dbNSFP (Liu, Wu, Li, & Boerwinkle, 2016), dbSNP or GnomAD (Kobayashi et al., 2017), as well as in-house information about the pathogenicity of *TP53* variants (T Soussi in preparation). For variant classification, we adapted the guidelines proposed by the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (Richards et al., 2015). Variants were classified as pathogenic (P), likely pathogenic (LP), possibly pathogenic (PP), likely benign (LB), and uncertain significance (VUS) based on the multiple predictive parameters incuded from dbNSFP as well as transcriptional activity based on a yeast-based functional assay included in the UMD_TP53 database.

The latest version of the UMD_TP53 database is available at https://p53.fr, a totally redesigned and updated website using the Joomla content management system. Due to database actualization, slight variations can be observed in the number of files included in the database and in Seshat.

Seshat

The block diagram describing data flow is shown in **Supp. Figure S2.** On the client side, Bootstrap (3.3.6) (http://getbootstrap.com) was used to design a layout compatible with all types of devices, mobile phones, tablets or desktop computers. Bootstrap elements were

styled with customized Bootswatch (3.3.7) (https://github.com/thomaspark/bootswatch) Darkly theme. jQuery (2.2.3) was used for JavaScript libraries (https://jquery.com/) mainly to implement client-side form validation and Odometer (0.4.6) (https://github.com/HubSpot/odometer) was used to create an animation of transitioning numbers on the front page.

On the server side, the application was developed on a Flask micro web framework (0.11.1) (http://flask.pocoo.org/), using Python 2.7 programming language. Flask was extended with Flask-WTF (0.13.1) (https://flask-wtf.readthedocs.io/) library that allows form validation, Flask-Mail (0.9.1) (https://pythonhosted.org/Flask-Mail/) library that enables the application to send emails, and Flask-Assets (0.12) (https://github.com/miracle2k/flask-assets) library that unifies and minifies JavaScript and CSS resources to speed up web page loading. Flask was installed to run on Apache 2 HTTP server (https://httpd.apache.org/). Pandas Python package (0.19.1) (http://pandas.pydata.org/) was used to handle data structures and data analysis. Cron software utility was used to periodically delete unnecessary files (user uploaded files and generated export files) from the server. The scheduler deletes two-day-old files every second day. Communication with Mutalyzer Name Checker was carried out over SOAP web service using the suds-jurko (0.6) (https://bitbucket.org/jurko/suds) Python package. Varcode (0.5.15) (https://github.com/hammerlab/varcode) Python library was deployed for parsing VCF and MAF batch files. Finally, xhtml2pdf (3.0.33) (https://github.com/xhtml2pdf/xhtml2pdf) Python library was used to create clinical export files, where HTML is converted into a PDF document.

As a starting point for annotations, Seshat uses the genomic position, reference and variant allele from MAF or VCF files. The size of these files is limited to 200 MB, allowing the analysis of several thousand variants. Seshat is able to extract *TP53* variants from among all other genes provided a correct genomic reference, based on hg18 (NCBI

Build 36.1), hg19 (GRCH37) or hg38 (GRCh38) genome builds, is used. CSV files using cDNA-based annotation (reference NM_000546.5) can also be used. A flexible input for individual analysis of single variants using genomic, cDNA or protein reference is also available. Using the HGVS recommendation, Seshat will transform this input using the stable NCBI sequence NG_017013.2 as a reference. Annotated variants are then analyzed using Mutalyzer (Wildeman et al., 2008). This procedure is straightforward for all SNV and no discrepancies should be detected between the annotation performed by Seshat and Mutalyzer provided no sequence errors are included in the input. This analysis is essential to keep all the data from Seshat according to the standards of HGVS-approved variant nomenclature (http://www.HGVS.org/varnomen/).

However, discrepancies may be observed for frameshift mutations, mostly due to inaccuracies in input files. We have observed two major problems. The first problem is related to small deletions (usually one or two nucleotides) that are not correctly assigned. The HGVS states that deletions should always be shifted to the 3′ and right-most position relative to the genomic sequence, a rule that is not always applied. This is partially due to the fact that the transcriptional orientation of the TP53 gene which is on the minus (-) strand as nucleotide numbering are based on the transcriptional orientation of the gene and goes in the opposite direction, creating confusing situations. Furthermore, as shown in **Supp. Figure S4A**, variants issued from short deletions can be described with a different nomenclature. Variant MN_000546.5:c.625_626delCT, the most frequent deletion oberved in the *TP53* gene is often described as MN_000546.5:c.624_625delTC as both events lead to the same final sequence. The second problem is associated with small insertions (generally comprising 1 to 5 nucleotides) that are in fact duplications **(Supp. Figure S4B).** We have observed that many pipelines do not handle duplications or inversions. For all of these problems, Seshat corrects the input variant and all subsequent analyses will be

performed using the correct variant nomenclature. Users will be notified of these changes in the final outputs.

Results are typically available by email within a few minutes depending on the number of *TP53* variants that have been analyzed. Each step of the process is closely monitored and error messages are displayed when a problem is encountered. In the final step, all variants are analyzed using data from the UMD_TP53 database. Seshat is not a repository database. The analysis is anonymous and batch files are not kept in the database. They are stored for a few days on our servers for debugging purposes before being automatically deleted.

The output files

Two output files are generated by Seshat, a condensed, short report with 73 fields and a long report with 164 different fields. The short output file contains essential information related to the variant and can be used as publication tables, whereas the long output file contains extensive information that can be useful for more detailed analysis. Both reports include descriptive and analytical information.

Descriptive information **(Table 1a)** is related to the correct annotation of the variant using multiple genomic, cDNA and protein references and versions. Names and versions of the data sources are always included in the header of output files. Analytical information **(Table 1b)** is related to *TP53* specific information associated with each variant, such as its frequency in the database, the residual *TP53* activity based on the work of Kato et al., as well as annotations from other databases (Kato et al., 2003). Well-characterized germline SNPs are also clearly identified. The long report also includes more statistical data issued from the analysis of the UMD_TP53 database as well as multiple information from the dbNSFP database version 3.5 (https://sites.google.com/site/jpopgen/dbNSFP) (Liu et al.,

2016). Both outputs also contain the input data from the original file. The various fields and examples of these two outputs are described in **Supp. Table S1, S2 and S3.** A complete documentation set is available to download (http://vps338341.ovh.net/help). It includes a read me file, a quick start document and samples files.

Implementation

As illustrated in **Figure 1**, in cancer genomics, analysing NGS sequencing data is a multistep process that typically involves multiple pipelines for data analysis: (i) bioinformatics tools for variant identification; (ii) variant annotation and prioritization; and (iii) interpretation of clinical significance by querying multiple databases. All these steps have been shown to be challenging and no gold standard pipeline is currently available.

Seshat was designed to circumvent all the problems associated with variant annotation as well as to provide the expert information brought by the LSDB UMD_TP53 (http://p53.fr/ and http://p53.fr/tp53-database/seshat).

Seshat **(Figure 3)** was primarily developed to automatically handle various types of data files (batch analysis). For NGS users, Seshat is compatible with both VCF and MAF files, the two most popular formats used to store NGS data. Seshat is able to retrieve *TP53* data from the bulk of genomic information with both file types, avoiding the need for any pre-treatment of the files to extract specific data. For Sanger sequencing, *TP53* mutations are often described using a cDNA-based nomenclature related to the transcript variant 1 (NM_000546). These variants can also be submitted to Seshat using CSV files. Input test files with different formats are available for download in the help section of the website. Individual analysis searching for a single variant can also be performed **(Figure 3)**.

Seshat was developed with the following specifications: i) simple input format requiring only position, reference and variant nucleotide ii) ensure minimal manual intervention to prevent typographical errors during sequence manipulation and iii) design a simple and comprehensive graphical interface (**Figure 4 and Supp. Figure S3A and**

S3B). For batch analysis, users can upload their data in a single step. Two output files in tab-separated values (TSV) format are generated within an hour of submission and are emailed to the user. The files are called ‘short’ and ‘long’ in reference to the number of variables that they include. The short file includes 73 variables, which are the most relevant out of the comprehensive list of 164 variables included in the long file **(**see Material and methods for the description of the variables included in each file**).** Both files contain analytical data as well as the input data to allow the user to monitor any specific issues that may have occurred during the process. Typical output files generated by Seshat are available in **Supp. Table S1, S2 and S3** and are partially described in **Table 1.**

In the first step, minimal genomic information such as genomic coordinates and genetic events is extracted to define a correct annotation using HGVS recommendations **(Figure 3)**. In the second step, the variant annotation is validated using the Name Checker tool developed by Mutalyzer (https://mutalyzer.nl/). Mutalyzer handles all types of variations that can target the *TP53* gene, such as substitutions, insertions, duplications, deletions, or more complex insertion/deletion (Wildeman et al., 2008). The current version of Mutalyzer (Mutalyzer 2.0.26) uses the stable NCBI sequence NG_017013.2 as a reference for *TP53* which is also used by LRG. This is a key issue, as it avoids any problems associated with the use of multiple genome references (HG18, HG19 or H38) by the various NGS pipelines. In a third step, the mutational data is compared to the UMD_TP53 database. Finally, in the last step, the information is displayed in tables and delivered to the user.

Table 1 presents a partial view of the central feature derived from Seshat for three representative variants. The upper part of the table **(** shows the description of the variant using genomic, cDNA and protein references. A full description of the reference in the title field avoids any possible ambiguities associated with the description. The current version of Seshat handles descriptions of a *TP53* variant for the 8 transcripts and 12 proteins

currently available in the stable Locus Reference Genomic sequence LRG_321 (http://ftp.ebi.ac.uk/pub/ databases/lrgex/LRG_321.xml).

The lower part of the table displays TP53-specific information derived from the UMD_TP53 database, such as frequency in the database, functional impact, and pathogenicity data (**see Material and methods for more information)**

A full description of each field is presented in the Seshat documentation available as **Supplementary material**.

The use of highly curated LSDB by experts in a specific domain, allows more accurate appraisal of multiple variants that cannot be performed by global analysis, as shown in the following two examples. Variants NM_000546.5:c.375G>A and NM_000546.5:c.375C>T both lead to the same synonymous protein variant (LRG_321p1:p.T125=) that is found 102 times in the database, both as somatic and germline variants (**see Table 1** for the analysis of this variant using Seshat**).** This nucleotide is located at the end of exon 4, just before the donor site in intron 4, and has been repeatedly shown to impair *TP53* splicing (Varley et al., 1998). This mutation is the most frequent synonymous SNV in the *TP53* database (Holmila, Fouquet, Cadranel, Zalcman, & Soussi, 2003). In several databases, this synonymous SNV is either not included or is defined as a benign variant despite our current knowledge of its pathogenicity (Soussi et al., 2017).

On the other hand, variant MN_000546.5:378C>G gives rise to a stop codon (LRG_321p1:p.T126*), which is usually defined as pathogenic. More detailed analysis of the consequence of this genetic event shows that it leads to a small shift in the splicing of the *TP53* gene and the synthesis of a full-length protein that only lacks a single amino acid at position 126 (Makarov et al., 2017). Functional analysis of this variant shows that it is indistinguishable from wild-type TP53.

These two examples show that many variants of a specific gene have their own specificity that must be taken into account. In the context of clinical analysis, particularly when

dealing with germline variants, it is essential to carefully validate all parameters before reporting clinical consequences of a given variant.

Discussion and future prospects

The clinical utility of *TP53* gene analysis is indisputable (Leroy et al., 2017). In CLL, analysis of *TP53* aberrations has been incorporated into routine clinical diagnostics to improve patient stratification and optimize therapeutic decisions (Malcikova et al., 2018). Other cancers such as AML will also benefit from an accurate *TP53* status analysis (Döhner et al., 2017). The *TP53* gene is mutated in more than 50% of human cancers reaching nearly 100% for high-grade serous ovarian carcinoma and small cell lung cancer (Soussi & Wiman, 2015). Thanks to the rapid pace of NGS sensitivity increase, these mutations could potentially be used as biomarkers for early detection in high-risk individuals, intermediate endpoints during treatment or for monitoring of disease recurrence (Phallen et al., 2017).

Transformation of high-throughput sequence variation descriptions found in VCF or MAF files into accurate nomenclature using the HGVS nomenclature is mandatory to ensure reliable interpretation of found variants. Although multiple commercial and non-commercial packages are available either as web services or standalone applications, all are generic and provide minimal specific information.

In contrast, LSDB benefits from rigorous expert curation, often coordinated by collaborating researchers with scientific expertise, but the formats of this database are highly heterogeneous and database maintenance is unpredictable (Soussi, 2014); Auerbach et al., 2011, #35201}.

Seshat, specifically developed for the analysis of *TP53*, combines unambiguous annotations as well as *TP53*-specific information in order to provide the most accurate picture of each novel or previously identified *TP53* variant.

Seshat is freely available from the *TP53* website (http://p53.fr/). Plans for further development include the possibility for users to perform statistical analysis of their dataset,

store their data in a private database and develop graphical outputs for enhanced display of the variants.

Acknowledgments

We are gratefull to Anna Piskorz, and Fanny Baran Marszak for beta testing and Antti Ylipaa for helpful discussions.

Conflict of Interest Statement

We have read and understood Human Mutation policy on declaration of interests and declare that we have no competing interests.

References

Ballinger, M. L., Mitchell, G., & Thomas, D. M. (2015). Surveillance recommendations for patients with germline TP53 mutations. *Curr Opin Oncol*, *27*(4), 332-337. doi:10.1097/CCO.0000000000000200

Bykov, V. J., & Wiman, K. G. (2014). Mutant p53 reactivation by small molecules makes its way to the clinic. *FEBS Lett*, *588*(16), 2622-2627. doi:10.1016/j.febslet.2014.04.017

Chen, M., & Manley, J. L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol*, *10*(11), 741-754. doi:10.1038/nrm2777

Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F. R., Büchner, T., . . . Bloomfield, C. D. (2017). Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*, *129*(4), 424-447. doi:10.1182/blood-2016-08-733196

Edlund, K., Larsson, O., Ameur, A., Bunikis, I., Gyllensten, U., Leroy, B., . . . Soussi, T. (2012). Data-driven unbiased curation of the TP53 tumor suppressor gene mutation database and validation by ultradeep sequencing of human tumors. *Proc Natl Acad Sci U S A*, *109*(24), 9551-9556. doi:10.1073/pnas.1200019109

Fernandez-Cuesta, L., Perdomo, S., Avogbe, P. H., Leblay, N., Delhomme, T. M., Gaborieau, V., . . . Brennan, P. (2016). Identification of Circulating Tumor DNA for the Early Detection of Small-cell Lung Cancer. *EBioMedicine*, *10*, 117-123. doi:10.1016/j.ebiom.2016.06.032

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, *17*(6), 333-351. doi:10.1038/nrg.2016.49

Holmila, R., Fouquet, C., Cadranel, J., Zalcman, G., & Soussi, T. (2003). Splice mutations in the p53 gene: case report and review of the literature. *Hum Mutat*, *21*(1), 101-102.

Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12497643

Joruiz, S. M., & Bourdon, J. C. (2016). p53 Isoforms: Key Regulators of the Cell Fate Decision. *Cold Spring Harb Perspect Med*, *6*(8). doi:10.1101/cshperspect.a026039

Kato, S., Han, S. Y., Liu, W., Otsuka, K., Shibata, H., Kanamaru, R., & Ishioka, C. (2003). Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc Natl Acad Sci U S A*, *100*(14), 8424-8429. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12826609

Khoo, K. H., Hoe, K. K., Verma, C. S., & Lane, D. P. (2014). Drugging the p53 pathway: understanding the route to clinical efficacy. *Nat Rev Drug Discov*, *13*(3), 217-236. doi:10.1038/nrd4236

Khoury, M. P., & Bourdon, J. C. (2010). The isoforms of the p53 protein. *Cold Spring Harb Perspect Biol*, *2*(3), a000927. doi:10.1101/cshperspect.a000927

Kobayashi, Y., Yang, S., Nykamp, K., Garcia, J., Lincoln, S. E., & Topper, S. E. (2017). Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. *Genome Med*, *9*(1), 13. doi:10.1186/s13073-017-0403-7

Krimmel, J. D., Schmitt, M. W., Harrell, M. I., Agnew, K. J., Kennedy, S. R., Emond, M. J., . . . Risques, R. A. (2016). Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. *Proc Natl Acad Sci U S A*, *113*(21), 6005-6010. doi:10.1073/pnas.1601311113

Leroy, B., Anderson, M., & Soussi, T. (2014). TP53 mutations in human cancer: database reassessment and prospects for the next decade. *Hum Mutat*, *35*(6), 672-688.

doi:10.1002/humu.22552

Leroy, B., Ballinger, M. L., Baran-Marszak, F., Bond, G. L., Braithwaite, A., Concin, N., . . . Soussi, T. (2017). Recommended Guidelines for Validation, Quality Control, and Reporting of TP53 Variants in Clinical Practice. *Cancer Res*, *6*, 1250-1260. doi:10.1158/0008-5472.CAN-16-2179

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat*, *37*(3), 235-241. doi:10.1002/humu.22932

Makarov, E. M., Shtam, T. A., Kovalev, R. A., Pantina, R. A., Varfolomeeva, E. Y., & Filatov, M. V. (2017). The rare nonsense mutation in p53 triggers alternative splicing to produce a protein capable of inducing apoptosis. *PLoS One*, *12*(9), e0185126. doi:10.1371/journal.pone.0185126

Malcikova, J., Tausch, E., Rossi, D., Sutton, L. A., Soussi, T., Zenz, T., . . . European, R. I. O. C. L. L. E. R. I. C. N. (2018). ERIC recommendations for TP53 mutation analysis in chronic lymphocytic leukemia-update on methodological approaches and results interpretation. *Leukemia*. doi:10.1038/s41375-017-0007-7

Merkle, F. T., Ghosh, S., Kamitaki, N., Mitchell, J., Avior, Y., Mello, C., . . . Eggan, K. (2017). Human pluripotent stem cells recurrently acquire and expand dominant negative P53 mutations. *Nature*, *545*(7653), 229-233. doi:10.1038/nature22312

Phallen, J., Sausen, M., Adleff, V., Leal, A., Hruban, C., White, J., . . . Velculescu, V. E. (2017). Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med*, *9*(403). doi:10.1126/scitranslmed.aan2415

Pospisilova, S., Gonzalez, D., Malcikova, J., Trbusek, M., Rossi, D., Kater, A. P., . . . Zenz, T. (2012). ERIC recommendations on TP53 mutation analysis in chronic lymphocytic leukemia. *Leukemia*, *26*(7), 1458-1461. doi:10.1038/leu.2012.25

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., . . . Committee, A. C.

M. G. L. Q. A. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*, *17*(5), 405-424. doi:10.1038/gim.2015.30

Soussi, T. (2014). Locus-specific databases in cancer: what future in a post-genomic era? The TP53 LSDB paradigm. *Hum Mutat*, *35*(6), 643-653. doi:10.1002/humu.22518

Soussi, T., Leroy, B., & Taschner, P. E. (2014). Recommendations for analyzing and reporting TP53 gene variants in the high-throughput sequencing era. *Hum Mutat*, *35*(6), 766-778. doi:10.1002/humu.22561

Soussi, T., Taschner, P. E., & Samuels, Y. (2017). Synonymous Somatic Variants in Human Cancer Are Not Infamous: A Plea for Full Disclosure in Databases and Publications. *Hum Mutat*, *38*(4), 339-342. doi:10.1002/humu.23163

Soussi, T., & Wiman, K. G. (2015). TP53: an oncogene in disguise. *Cell Death Differ*, *22*(8), 1239-1249. doi:10.1038/cdd.2015.53

Taschner, P. E., & den Dunnen, J. T. (2011). Describing structural changes by extending HGVS sequence variation nomenclature. *Hum Mutat*, *32*(5), 507-511. doi:10.1002/humu.21427

Trounson, A. (2017). Potential Pitfall of Pluripotent Stem Cells. *N Engl J Med*, *377*(5), 490-491. doi:10.1056/NEJMcibr1706906

Varley, J. M., McGown, G., Thorncroft, M., White, G. R., Tricker, K. J., Kelsey, A. M., . . . Evans, D. G. (1998). A novel TP53 splicing mutation in a Li-Fraumeni syndrome family: a patient with Wilms' tumour is not a mutation carrier. *Br J Cancer*, *78*(8), 1081-1083. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9792154

Wildeman, M., van Ophuizen, E., den Dunnen, J. T., & Taschner, P. E. (2008). Improving

sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum Mutat*, *29*(1), 6-13. doi:10.1002/humu.20654

Yen, J. L., Garcia, S., Montana, A., Harris, J., Chervitz, S., Morra, M., . . . Church, D. M. (2017). A variant by any name: quantifying annotation discordance across tools and clinical databases. *Genome Med*, *9*(1), 7. doi:10.1186/s13073-016-0396-7

Figures Legends

Figure 1: Typical NGS data analysis pipeline for cancer genome sequencing. Once the sample has been sequenced, several steps are necessary to convert the complex raw data into meaningful information that can subsequently be used to query multiple databases. Erroneous annotations can interfere with this last analytical step. Seshat provides (1) correct annotation for any *TP53* variant, (2) *TP53* information obtained from multiple databases and unique statistical analysis derived from the UMD_*TP53* database.

Figure 2: Description of the 2017 release of the UMD_*TP53* database. Among the 80,406 *TP53* variants reported in the literature (right panel), 6,874 different *TP53* variants have been identified (left panel). Although frameshift variants appear to be more frequent, they are not found repeatedly in the database compared to missense variants with several hot spots that are very frequent (see text for more information)(Leroy, Anderson, & Soussi, 2014; Soussi et al., 2014).

Figure 3: Overview of the Seshat pipeline. Sequencing data are converted to a homogeneous format based on genomic nomenclature and tested by Mutalyzer web tools (left panel). *TP53* specific information is then retrieved from the UMD_TP53 database to build output tables (right panel).

Figure 4: Screenshot of the web-based Graphical User Interface of Seshat displaying the manual submission panel (see **Supp. Figure S4A and S4B f**or a detail description**)**