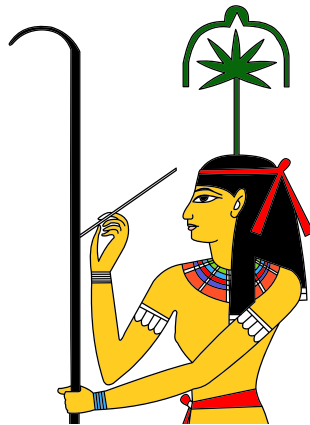


Seshat:

An innovative tool to handle somatic and germline TP53 variants



Read me document

Version 1.0 October 2017

Table of Content

Seshat: an innovative tool to handle somatic and germline TP53 variants	3
TP53 gene organization and numbering.	5
Seshat: an overview.	6
How to use Seshat.....	7
A quick start	7
Output tables.....	7
Seshat description: Home page general.....	8
Batch analysis.....	9
File Format.....	10
Seshat: Single analysis	12
Single analysis: DNA/RNA: Single Nucleotide Variant	12
Output screen: SNV.....	13
Single analysis: DNA/RNA: Deletion.....	14
Output screen: deletion.....	15
Single analysis: DNA/RNA: Insertion	16
Output screen insertion example 1	17
Output screen insertion: example 2.....	18
Output screen insertion: example 3.....	18
Single analysis: DNA/RNA: Insertion + Deletion.....	19
Output screen insertion/deletion	20
Single analysis: DNA/RNA: duplication.....	21
Output screen duplication.....	22
Q&A	23

Comments ? problems ? one address : p53@free.fr

Seshat: an innovative tool to handle somatic and germline TP53 variants



Seshat was the Ancient Egyptian goddess of wisdom, knowledge, and writing. Her name means she who is the scribe, which is in perfect harmony with the goal of this site.*

*Adapted from <https://en.wikipedia.org/wiki/Seshat>

Seshat performs the following tasks:

- Quality check mutation nomenclature.

- Generates a full description of each variant formatted according to hgvs.

- Generates publication-ready tables.

- Assesses the pathogenicity of each variant according to either general prediction algorithms (Provean, Sift, Polyphen2, FATHMM, MutationAssessor and 7 other algorithms) or algorithms developed exclusively for TP53.

- Displays functional and structural data for each TP53 variant.

Seshat works with CSV files generated by the user and VCF or MAF files generated by NGS. Input data using protein, cDNA or genomic information can be used.

Seshat is based on the UMD TP53 database; 2017 release: 68,000 mutations (7,000 TP53 variants).

Introduction

Somatic mutations in the TP53 gene are one of the most frequent alterations in human cancers, and germline mutations are the underlying cause of Li-Fraumeni syndrome, which predisposes to a wide spectrum of early-onset cancers.

Accurate assessment of TP53 gene status in sporadic tumors and in the germline of individuals at high risk of cancer has important clinical implications for diagnosis, surveillance and therapy.

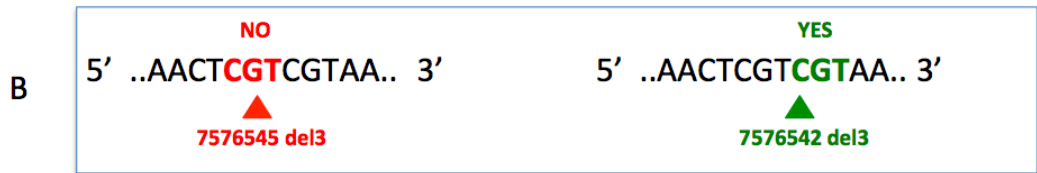
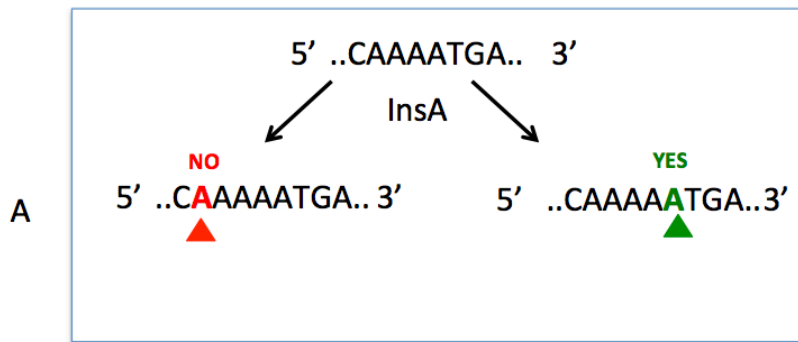
The most recent version of the National Comprehensive Cancer Network (NCCN) guidelines recommends TP53 mutations testing in individuals with onset of breast cancer before 31 years of age, either concurrently with BRCA1/2 testing or as a follow-up test after negative BRCA1/2 testing (NCCN Guidelines Version 1.2017, http://www.nccn.org/professionals/physician_gls/pdf/genetics_screening.pdf).

Somatic TP53 mutation analysis is now widely used in clinical trials involving patient stratification based on TP53 status and in trials of novel drugs targeting either wild-type or mutant TP53 in order to activate a TP53 antitumor response. TP53 mutation screening is therefore rapidly becoming an integral part of many therapeutic or prevention strategies in clinical practice.

The complex architecture and expression pattern of the TP53 gene has only been recognized in recent years. TP53 mobilizes various mechanisms to transcribe at least eight different mRNA isoforms, which are generated by alternative splicing or alternative promoter usage. Collectively, these mRNAs have the potential to give rise to up to 12 different proteins, although the exact expression level, tissue distribution and biological function of each of these protein variants are poorly understood. This complex expression pattern implies that sequences located in TP53 introns and involved in the production of alternative forms of the protein may have a critical impact on overall biological functions of p53, and may therefore be important target regions for somatic or germline variants.

Bioinformatic pipelines currently associated with various NGS devices use different nomenclatures, algorithms and references to assess TP53 status, leading to heterogeneous outputs. For example, we have observed that variant chr17:g.7578406G>A (MN_000546.5_c.524G>A) can be randomly described as either p.R175H (using NP_000537.3 as a reference) or p.R43H (using NP_001119587.1 as a reference) in the same output. The fact that no reference is usually included results in the misleading conclusion that two different TP53 variants have been identified.

Another problem concerns the mutation nomenclature, which is not homogeneous among the various outputs and rarely follows the hgvs recommendations (<http://varnomen.hgvs.org/>), leading to the false impression that a given mutation is a novel mutation (Figure). The 3' rule used for the description of mutations in repeated sequences is rarely followed.



The 3' rule, a misapplied rule:

For all descriptions, the most 3' position possible of the reference sequence is arbitrarily assigned to have been changed.

The 3' rule also applies to changes in single residue stretches and tandem repeats (nucleotide or amino acid).

The 3' rule applies to ALL descriptions (genome, gene, transcript and protein) of a given variant.

More info here:

<http://varnomen.hgvs.org/recommendations/general/>

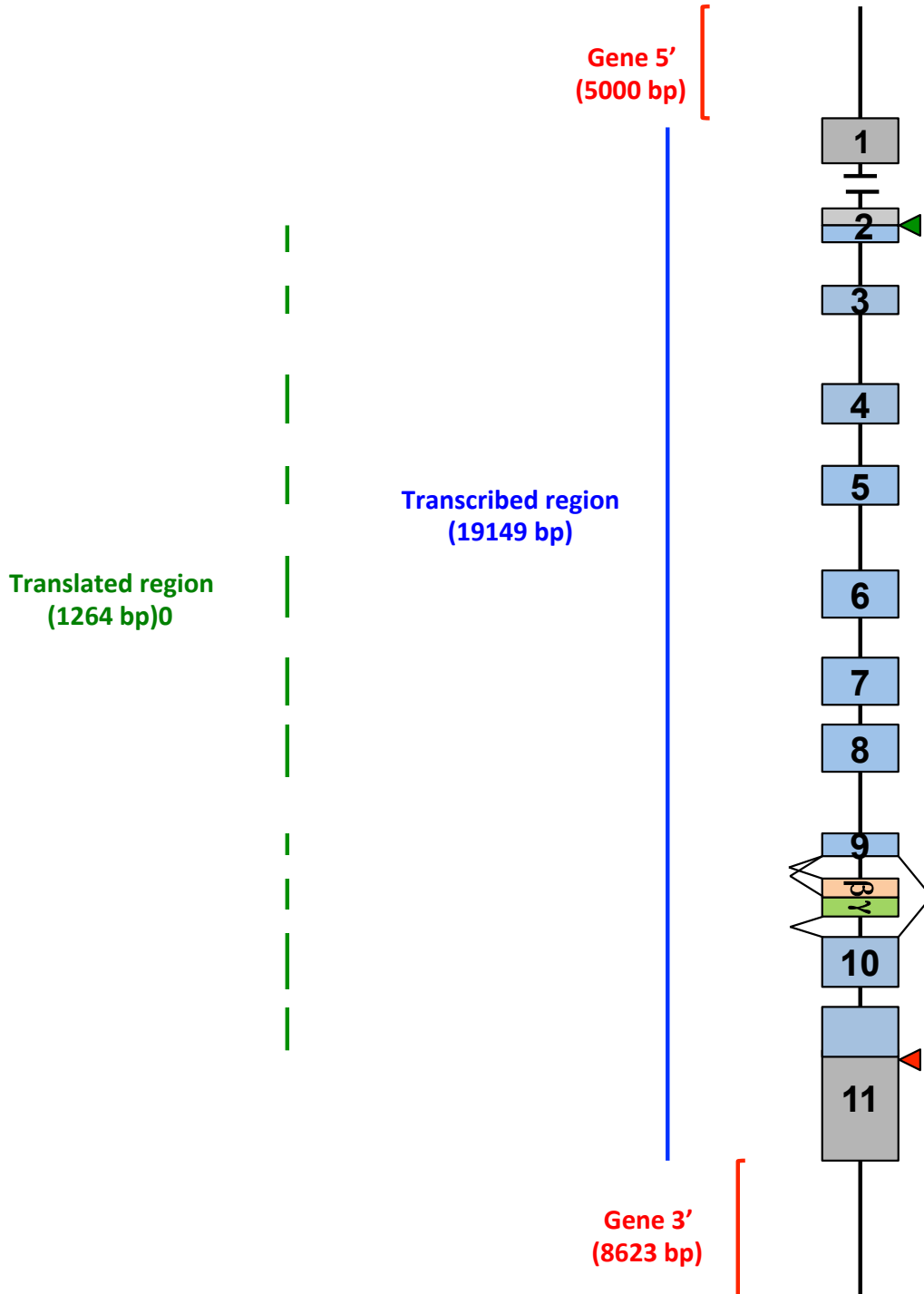
A: Insertion of nucleotide A indifferently at various positions leads to multiple mutational events despite a similar outcome.

B: Deletion of multiple nucleotides can lead to the same final sequence. The most 3' event leading to the final variant must be used to generate a unique and reproducible mutational event.

Seshat was developed to resolve all of these problems. Using raw data (VCF, MAF or CSV files), Seshat generates multiple output tables with accurate and complete information on each TP53 variant.

TP53 gene organization and numbering.

	HG18	HG19	HG38	NG_01713.2
Start	7536593	7595868	7692550	1
End	7503822	7563097	7659779	32772

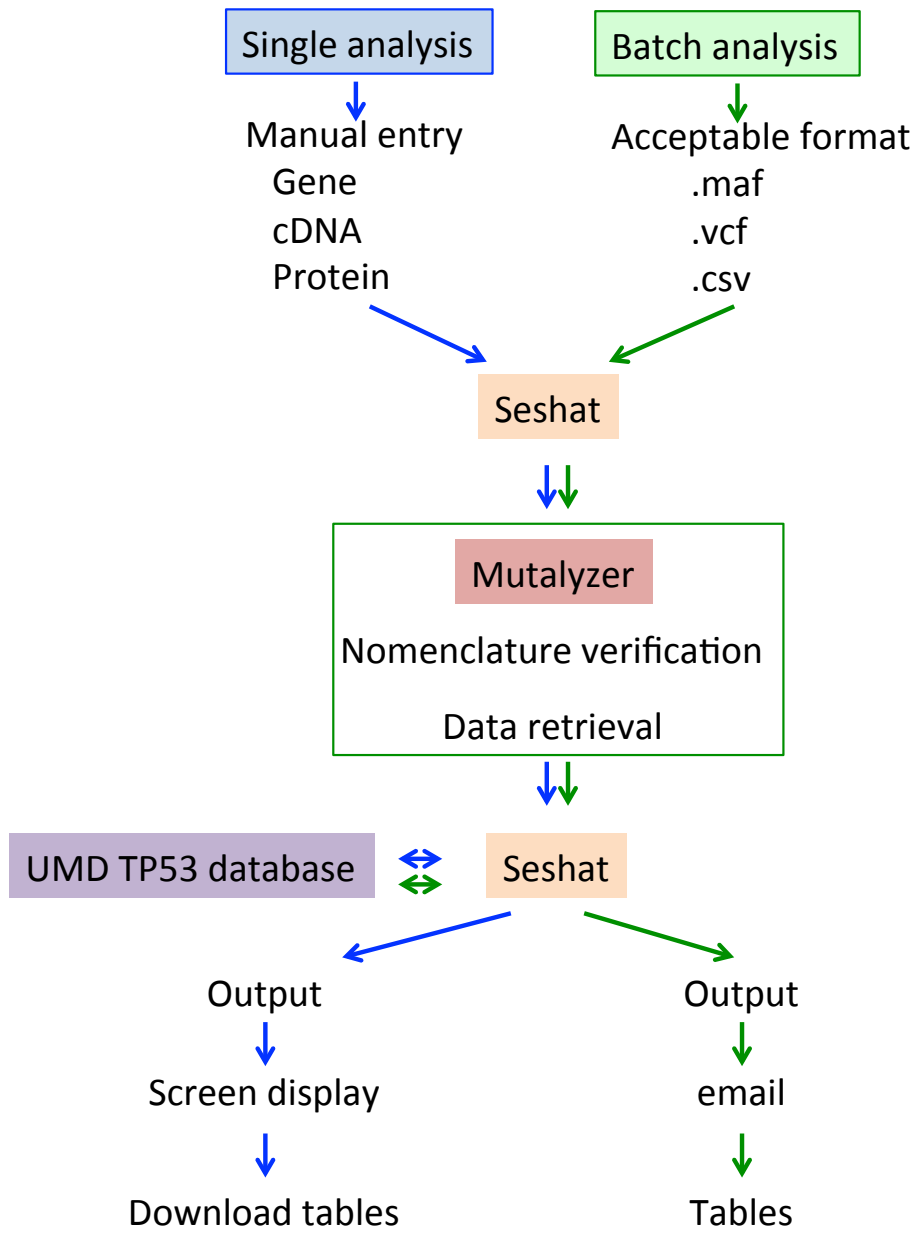


Seshat will analyze the entire TP53 gene using NG_01713.2, the most updated reference of the gene. It includes all the coding regions as well as part of the 5' and 3' region.

(https://www.ncbi.nlm.nih.gov/nucore/NG_017013.2).

The three genome nomenclature (HG18, HG19 and HG39) can be used for the analysis. Beware that the numbering from 5' to 3' is decreasing.

Seshat: an overview.



How to use Seshat

A quick start

Batch analysis

- 1- download the test files (MAF, VCF or CSV) on your computer
- 2- go to the “batch analysis” page
- 3- upload one of the test files using the “select file” button”
- 4- fill your email address in the dialog box
- 5- for VCF files, please indicate the genome build version used in your analysis.
- 6- click the button “start the analysis”

You will receive an email with the result of the analysis within one hour.

Single analysis

- 1- go to the “single analysis” page
- 2- fill the dialog box “Start position” with “7577538”
- 3- fill the dialog box “Mutant allele” with “A”
- 4- click the button “start the analysis”

A display of the results is shown on the screen

Click the button “export table” if you want a full description of the mutant

Output tables

Two output files in tab-separated values (TSV) format are generated within the hour of the submission and will be emailed to the user. The short output contains essential information related to the variant and can be used as publication tables whereas the long output contains extensive information that can be useful for deeper analysis. The various fields and examples of these two outputs are described in the annex. Both outputs also contain the input data from the original file.

A full description of the various items described in the output files is available in the two PDF documents that be be downloaded in the website.

Seshat description: Home page general

The screenshot shows the Seshat website interface. At the top, there are navigation links: Seshat (1), Single analysis (2), Batch analysis (3), TP53 website (4), Help (5), and About (6). The main header features the Seshat logo and a description: 'An innovative tool to handle somatic and germline TP53 variants.' Below this, it lists tasks performed by Seshat, such as quality checks, generating descriptions, and assessing pathogenicity. A statistics bar shows 5,650 Variants (7), 63,000 Mutations, and 4,100 Publications (8). The main form area includes fields for Reference sequence (1), Variant type (2), Start position (3), End position (4), Wild type allele (5), Mutant allele (6), Strand polarity (7), and Sample ID (8). A 'Start the analysis' button (9) is at the bottom.

1	Seshat home page	6	Link to the info page
2	Link to Seshat manual analysis	7	Tab linked to the DNA/RNA analysis
3	Link to Seshat batch analysis	8	Tab linked to the protein analysis
4	Link to the TP53 website		
5	Link to the Seshat help page	9	Button start once all the fields have been filled.

Input Fields for a manual analysis (1 variant)

1	Reference sequence	Versions of human genome used for the analysis
2	Variant type	SNV (Single Nucleotide Variant), deletion, insertion duplication or indel (insertion + deletion)
3	Start position	Position of mutation start using the reference defined in box 1
4	End position	Position of mutation end using the reference defined in box 1
5	Wild type allele	Wild type nucleotide a position start
6	Mutant allele	Mutant nucleotide(s)
7	Strand polarity	Self-explanatory
8	Sample ID	Self-explanatory

Batch analysis

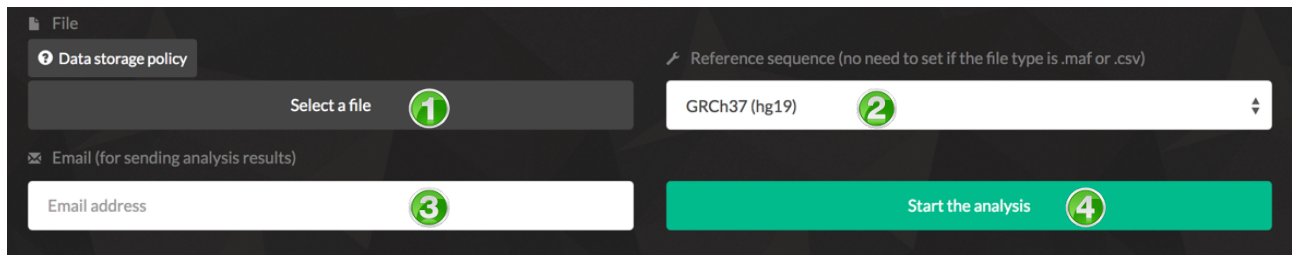
A quick, efficient and accurate way to analyze and review multiple TP53 variants.
For NGS users, VCF or MAF files can be used.

For variants derived from conventional sequencing, only data using the cDNA reference **NM_000546.5** can be used.

The size of the file must not exceed 200 Mo, corresponding to several thousand mutations.

Important. For VCF and MAF files, it is unnecessary to upload files that contain only TP53 mutations. Seshat can handle files that contain variants in multiple genes and will extract data specific for TP53.

For your convenience, test files (MAF, VCF or CSV) are available for download to check the program.



The screenshot shows a dark-themed web interface for batch analysis. At the top left, there is a 'File' menu and a 'Data storage policy' link. Below this is a 'Select a file' button with a green circle containing the number 1. To the right is a 'Reference sequence' dropdown menu with 'GRCh37 (hg19)' selected and a green circle containing the number 2. Below the file selection is an 'Email (for sending analysis results)' section with an 'Email address' input field and a green circle containing the number 3. At the bottom right is a large green 'Start the analysis' button with a green circle containing the number 4.

Step 1: green stamp 1
Select the file to be uploaded.

Step 2: green stamp 2
Choose the reference sequence for VCF file
(not necessary for MAF and CSV file)

Step 3: green stamp 3
Provide an email address.
By default, the name of the uploaded file will be used for the output file.

Step 4
Start the analysis.

An email will be send with the output tables
With the test files, results will be available withing 1 to 5 minutes

For bigger files with more than 1 000 TP53 mutations the processing time will be longer and can take one hour. The processing time is related to the number of TP53 mutations, not the number of total mutation.

Results of batch analysis - TCGA.LUSC.mutect.74e193ff-2f2c-4722-b12d-d3507626ee00.somatic.maf

Analyzed batch file: TCGA.LUSC.mutect.74e193ff-2f2c-4722-b12d-d3507626ee00.somatic.maf
Time taken to run the analysis: 3 minutes 5 seconds
The input file contained 172813 mutations out of which 62 were TP53 mutations.

example: a 190 Mo file with 172813 mutations.

File Format

MAF and VCF

MAF or VCF files will work as long as they use the official standard format.



We have noticed that there are some heterogeneity in MAF and VCF file headers that can lead to some problems with Seshat. If your file is not handled properly by Seshat, please check the format of the heading columns in the example file.

VCF specification

VCF is a text file format. It contains

- i) meta-information lines,
- ii) a header line,
- iii) data lines each containing information about a position in the genome as well as various information regarding the variant

If you have multiple VCF files (one file for each patient), it is possible to concatenate these files with galaxy (<https://usegalaxy.org>) using the VCF combine script.

VCF files generated by galaxy are fully compatible with Seshat.



never ever modify VCF files with Excel. Use dedicated softwares or textfile editors.

The VCF test file in the Seshat website will generate an email with this message:

Results of batch analysis

Analyzed batch file: VCF_test_Seshat.vcf

Time taken to run the analysis: 0 minutes 8 seconds

Summary: The input file contained 21 mutations out of which 14 were TP53 mutations.

The MAF test file in the Seshat website will generate an email with this message:

Results of batch analysis

Analyzed batch file: Test.maf

Time taken to run the analysis: 0 minutes 16 seconds

Summary: The input file contained 5074 mutations out of which 22 were TP53 mutations.

CSV file

Use the the reference **NM_000546.5** to described the variant

Ex:

Mutation, Patient ID

c.743G>A, test1

c.626_627del2, test2

c.559+1G>A, test3

c.741_742delCCInsTT, test4

c.390_426del37, test5

c.-28-2671_-28-2666del6, test6

c.604_605insTAT, test7

c.1041_1042insGAGAGCTGAATGAGGCC, test8

The CSV test file in the Seshat website will generate an email with this message:

Results of batch analysis

Analyzed batch file: Seshat_test_csv.csv

Time taken to run the analysis: 0 minutes 31 seconds

Summary: The input file contained 58 mutations out of which 58 were TP53 mutations.

Seshat: Single analysis

Single analysis: DNA/RNA: Single Nucleotide Variant

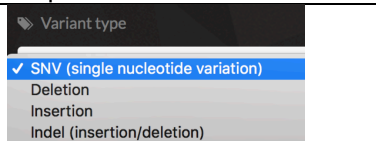
Data example.

Reference sequence	Variant type	Start position	End position	Wild-type allele	Mutant allele	Strand polarity	Sample ID*
hg19	SNV	7577538	7577538	G	T	Positive	test
hg18	SNV	7520037	7520037	G	A	Positive	testb
NM_000546.5	SNV	524	525	G	A	Positive	testc

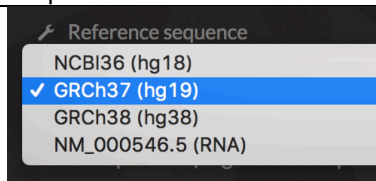
Information generated by the program that cannot be modified is shown in red.

* optional

Step 1:

	Choose the type of variant: SNV
---	---------------------------------

Step 2:

	Choose the reference sequence (hg18, hg19, hg38) for genomic entry or the full-length RNA NM_000546.5 for cDNA entry.
---	---

A full description of the TP53 gene and the various TP53 transcripts can be found at the Ensembl website:

(http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=LRG_321;r=17:7668402-7687550)

or at the LRG website

(http://ftp.ebi.ac.uk/pub/databases/lrgex/LRG_321.xml)

Step 3:

GRCh37 (hg19)	SNV (single nucleotide variation)
Start position (larger than end position)	End position
7577538	7577538
Wild type allele (coding strand)	Mutant allele (coding strand)
G	e.g., ATGC...

Choose the start position. The wild-type allele will be displayed automatically.

Any value outside the range of the various references will generate an error message.

Step 4:

GRCh37 (hg19)	SNV (single nucleotide variation)
Start position (larger than end position)	End position
7577538	7577538
Wild type allele (coding strand)	Mutant allele (coding strand)
G	A

Enter the variant allele. An error message will be generated if wt. and mutant alleles are similar. Input is limited to a single nucleotide (G, A, T or C).

Step 5: (Optional)

Most mutations are described using the positive strand of the gene (**set by default**). If the mutation is described in the negative strand, change the value here.

Step 7 (Optional)
Step 8

An ID can be given to the analysis.
Start the analysis.

Output screen: SNV

1	Input data	Summary of the user data.
2	Nomenclature	Official nomenclature of the variant using the three references of the human genome.
3	Description	Important features of the variant. Each item is fully described in the annex document.
4	Full description according to LRG	Full description of the variant in each TP53 transcript according to LRG.
5		Full description of the variant for each TP53 protein isoform according to LRG.
6	TP53 proteins	Description of the variant for the major TP53 protein isoforms; localization of the variant in the various domains of the protein.
7	Comments	Specific comments regarding the TP53 variant. Each item is fully described in the annex document.

10	Summary export	Generate a pdf file with an extensive description of the TP53 variants.
11	Export tables	Generate two TSV files with a full analysis of the variant: see accompanying documents for more information.
12	Back to homepage	Self-explanatory.

Single analysis: DNA/RNA: deletion (max size of a deletion handled by Seshat: 200 bp)*

Data example.

Reference sequence	Variant type	Start position	End position**	Wild-type allele	Mutant allele	Strand polarity	Sample ID***
hg19	Deletion	7578223	7578222	GA		Positive	test
NM_000546.5	Deletion	524	527	GCTG		Positive	test

Information generated by the program that cannot be modified is shown in red.

* a note of caution: deletion sizes are calculated using the genomic reference. The size of a deletion in the cDNA can be different in the gene if it spans two different exons.

** For genome position start and end correspond to the 5' and 3' boundaries of the deletion respectively and only include deleted nucleotides. As the numbering of the TP53 gene from 5' to 3' is decreasing, value of start position is always higher than end position for frameshift mutations.

*** optional

Step 1:

<input type="checkbox"/> SNV (single nucleotide variation) <input checked="" type="checkbox"/> Deletion <input type="checkbox"/> Insertion <input type="checkbox"/> Indel (insertion/deletion)	Choose the type of variant: Deletion
---	--------------------------------------

Step 2:

<input checked="" type="checkbox"/> Reference sequence <input type="checkbox"/> NCBI36 (hg18) <input checked="" type="checkbox"/> GRCh37 (hg19) <input type="checkbox"/> GRCh38 (hg38) <input type="checkbox"/> NM_000546.5 (RNA)	Choose the reference sequence (hg18, hg19, hg38) for genomic entry or the full-length RNA NM_000546.5 for cDNA entry.
---	---

A full description of the TP53 gene and the various TP53 transcripts can be found at the Ensembl website: (http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=LRG_321;r=17:7668402-7687550) or at the LRG website

(http://ftp.ebi.ac.uk/pub/databases/lrgex/LRG_321.xml)

Step 3 and 4:

GRCh37 (hg19)	Deletion
Start position (larger than end position)	End position
7578223	7578222
Wild type allele (coding strand)	Mutant allele (coding strand)
GA	e.g., ATGC...
Strand polarity (leave positive if you have no information)	Sample ID (optional)
Positive	Sample ID

Choose the start position.
 The wild-type allele will be displayed automatically.
 Any value outside the range of the various references will generate an error message.
 Enter the end position

Step 5: (Optional)

Most mutations are described using the positive strand of the gene (set by default). If the mutation is described in the negative strand, change the value here.

Step 7 (Optional)

An ID can be given to the analysis.

Step 8

Start the analysis.

9	Button start once all the fields have been filled.
---	--

Output screen: deletion

1	Input data	Summary of the user data.
2	Nomenclature	Official nomenclature of the variant using the three references of the human genome. When the mutation has not been detected in human cancer, only the HGVS format is used.
3	Description	When the mutation has not been described in the UMD database, this warning message will be displayed
4	Full description according to LRG	Full description of the variant in each TP53 transcript according to LRG.
5		Full description of the variant for each TP53 protein isoform according to LRG.
6	TP53 proteins	Description of the variant for the major TP53 protein isoforms; localization of the variant in the various domains of the protein.
7	Comments	Specific comments regarding the TP53 variant. Each item is fully described in the annex document.
10	Summary export	Generate a pdf file with an extensive description of the TP53 variants.
11	Export tables	Generate two TSV files with a full analysis of the variant: see accompanying documents for more information.
12	Back to homepage	Self-explanatory.

Single analysis: DNA/RNA: Insertion

Data example.

Reference sequence	Variant type	Start position	End position	Wild-type allele	Mutant allele	Strand polarity	Sample ID*
hg19	Insertion	7578529	7578528	TT	AT	Positive	test
hg19	Insertion	7578469	7578468	GC	GCCCGGC	positive	test
NM_000546.5	Insertion	626	627	TT	AT	positive	test

Information generated by the program that cannot be modified is shown in red.

* For genome position start and end correspond to the 5' and 3' boundaries of the deletion respectively and only include deleted nucleotides. As the numbering of the TP53 gene from 5' to 3' is decreasing, **value of start position is always higher than end position for frameshift mutations.**

This feature is valid for genomic reference only, not for RNA / cDNA.

** optional

Step 1:

<div style="border: 1px solid black; padding: 5px;"> <p>SNV (single nucleotide variation)</p> <p>Deletion</p> <p><input checked="" type="checkbox"/> Insertion</p> <p>Indel (insertion/deletion)</p> </div>	Choose the type of variant: Insertion
---	---------------------------------------

Step 2:

<div style="border: 1px solid black; padding: 5px;"> <p>Reference sequence</p> <p>NCBI36 (hg18)</p> <p><input checked="" type="checkbox"/> GRCh37 (hg19)</p> <p>GRCh38 (hg38)</p> <p>NM_000546.5 (RNA)</p> </div>	Choose the reference sequence (hg18, hg19, hg38) for genomic entry or the full-length RNA NM_000546.5 for cDNA entry.
---	---

A full description of the TP53 gene and the various TP53 transcripts can be found at the Ensembl website:

(http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=LRG_321;r=17:7668402-7687550)

or at the LRG website

(http://ftp.ebi.ac.uk/pub/databases/lrgex/LRG_321.xml)

Step 3 and 4:

GRCh37 (hg19)	Insertion
Start position (larger than end position)	End position
7578529	7578528
Wild type allele (coding strand)	Mutant allele (coding strand)
TT	TT

Step3

Choose the start position.

The wild-type allele will be displayed automatically.

Any value outside the range of the various references will generate an error message.

The end position will be set automatically

Step 4

Enter the insertion

Important:

In several instances, the inserted sequences are unknown. Seshat allows some flexibility for insertion.

If the length of the inserted sequence is known, it is possible to fill the mutant allele fields with either "NNNNN" or "5" if 5 nucleotides have been inserted (see example 2).

If the sequence is unknown, it is possible to use "?" but the output will not be able to display accurately the consequence of the variation (see example 3).

Step 5: (Optional)

Most mutations are described using the positive strand of the gene (set by default). If the mutation is described in the negative strand, change the value here.

Step 6 (Optional)

An ID can be given to the analysis.

Step 7

Start the analysis.

9	Button start once all the fields have been filled.
----------	--

Output screen insertion example 1

The screenshot shows the TP53 analysis web interface. At the top, there are navigation links for 'Seshat', 'Single analysis', 'Batch analysis', 'TP53 website', 'Help', and 'About'. Below this is the 'Input data' section with fields for Reference (HG19), Start (7578529), End (7578528), Wild type (TT), Mutant (AT), and Type (INS). The main content area is divided into several panels: 'Nomenclature' (2), 'Description' (3), 'Full description according to LRG' (4, 5), 'TP53 proteins' (6), and 'Comments' (7). At the bottom, there are three buttons: 'Clinical export' (10), 'Export tables' (11), and 'Back to homepage' (12).

1	Input data	Summary of the user data.
2	Nomenclature	Official nomenclature of the variant using the three references of the human genome.
3	Description	Important features of the variant. Each item is fully described in the annex document.
4	Full description according to LRG	Full description of the variant in each TP53 transcript according to LRG.
5		Full description of the variant for each TP53 protein isoform according to LRG.
6	TP53 proteins	Description of the variant for the major TP53 protein isoforms; localization of the variant in the various domains of the protein.
7	Comments	Specific comments regarding the TP53 variant. Each item is fully described in the annex document.

10	Summary export	Generate a pdf file with an extensive description of the TP53 variants.
11	Export tables	Generate two TSV files with a full analysis of the variant: see accompanying documents for more information.
12	Back to homepage	Self-explanatory.

Output screen insertion: example 2

Inserted sequence: NNNNN

Input data					
Reference	Start	End	Wild type	Mutant	Type
HG19	7578469	7578468	GC	NNNNN	INS

HGVS recommendations

Only the insertion of a sequence or a range is implemented.

Nomenclature

Mutation in HGVS format
NG_017013.2:g.17400_17401ins(NNNNN)

Clinical export
Export tables
Back to homepage

Output screen insertion: example 3

Inserted sequence: ?

Input data					
Reference	Start	End	Wild type	Mutant	Type
HG19	7579471	7579470	CG	?	INS

Nomenclature

Mutation in HGVS format
NG_017013.2:g.16398_16399ins(?)

NCBI36 (hg18)
chr17:g.7520196_7520195ins(?)

GRCh37 (hg19)
chr17:g.7579471_7579470ins(?)

GRCh38 (hg38)
chr17:g.7676153_7676152ins(?)

Full description according to LRG

Affected proteins

LRG_321p1:p.(P72fs)
LRG_321p3:p.(P72fs)
LRG_321p4:p.(P72fs)
LRG_321p5:p.(=)
LRG_321p6:p.(=)
LRG_321p7:p.(=)
LRG_321p8:p.(P33fs)
LRG_321p9:p.(P33fs)
LRG_321p10:p.(P33fs)
LRG_321p11:p.(=)
LRG_321p12:p.(=)
LRG_321p13:p.(=)

TP53 proteins

TP53 alpha
p.(P72fs)

TP53 beta
p.(P72fs)

TP53 gamma
p.(P72fs)

TP53 domain
Proline Rich

Structural motif
-

Post-translational modifications
-

Description

Records in database
2

Classification
Frame_Shift_Ins

Comment
Exonic mutation

Frequency
This frameshift variant is rare

HGVS recommendations

Expected "-" (at char 30), (line:1, col:31)

Comments

Activity
The activity of truncated p53 is assumed to be nil

Isoforms
-

Prediction
No prediction for frameshift mutation

Clinical export
Export tables
Back to homepage

Single analysis: DNA/RNA: Insertion + deletion

Data example.

Reference sequence	Variant type	Start position	End position	Wild-type allele	Mutant allele	Strand polarity	Sample ID*
hg19	Indel	7574003	757402	CC	TT	Positive	test

Information generated by the program that cannot be modified is shown in red.

* For genome position start and end correspond to the 5' and 3' boundaries of the deletion respectively and only include deleted nucleotides. As the numbering of the TP53 gene from 5' to 3' is decreasing, **value of start position is always higher than end position for frameshift mutations.**

** optional

Step 1:

SNV (single nucleotide variation) Deletion Insertion <input checked="" type="checkbox"/> Indel (insertion/deletion)	Choose the type of variant: Indel
--	-----------------------------------

Step 2:

Reference sequence NCBI36 (hg18) <input checked="" type="checkbox"/> GRCh37 (hg19) GRCh38 (hg38) NM_000546.5 (RNA)	Choose the reference sequence (hg18, hg19, hg38) for genomic entry or the full-length RNA NM_000546.5 for cDNA entry.
--	---

A full description of the TP53 gene and the various TP53 transcripts can be found at the Ensembl website: (http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=LRG_321;r=17:7668402-7687550) or at the LRG website (http://ftp.ebi.ac.uk/pub/databases/lrgex/LRG_321.xml)

Steps 3 and 4:

GRCh37 (hg19)	Indel (insertion/deletion)
Start position (larger than end position)	End position
7574003	7574002
Wild type allele (coding strand)	Mutant allele (coding strand)
CG	TT

Step3
Choose the start position.
The wild-type allele will be displayed automatically.
Any value outside the range of the various references will generate an error message.
The end position will be set automatically

Step 4 Choose the End position.

Step 5 Enter the insertion

Important:

In several instances, the inserted sequences are unknown. Seshat allows some flexibility for insertion.

If the length of the inserted sequence is known, it is possible to fill the mutant allele fields with either "NNNNN" or "5" if 5 nucleotides have been inserted.

If the sequence is unknown, it is possible to use "?" but the output will not be able to display accurately the consequence of the variation.

Step 6: (Optional)

Most mutations are described using the positive strand of the gene (set by default). If the mutation is described in the negative strand, change the value here.

Step 7 (Optional)

An ID can be given to the analysis.

Step 8

Start the analysis.

9	Button start once all the fields have been filled.
----------	--

Output screen insertion/deletion

1	Input data	Summary of the user data.
2	Nomenclature	Official nomenclature of the variant using the three references of the human genome.
3	Description	Important features of the variant. Each item is fully described in the annex document.
4	Full description according to LRG	Full description of the variant in each TP53 transcript according to LRG.
5		Full description of the variant for each TP53 protein isoform according to LRG.
6	TP53 proteins	Description of the variant for the major TP53 protein isoforms; localization of the variant in the various domains of the protein.
7	Comments	Specific comments regarding the TP53 variant. Each item is fully described in the annex document.

10	Summary export	Generate a pdf file with an extensive description of the TP53 variants.
11	Export tables	Generate two TSV files with a full analysis of the variant: see accompanying documents for more information.
12	Back to homepage	Self-explanatory.

Single analysis: DNA/RNA: duplication

Data example.

Reference sequence	Variant type	Start position	End position	Wild-type allele	Mutant allele	Strand polarity	Sample ID*
hg19	duplication	7579471	7579471	C	C	Positive	test
NM_000546.5	duplication	1001	1006	GGCGTG	GGCGTG		

Information generated by the program that cannot be modified is shown in red.

* For genome position start and end correspond to the 5' and 3' boundaries of the deletion respectively and only include deleted nucleotides. As the numbering of the TP53 gene from 5' to 3' is decreasing, **value of start position is always higher than end position for frameshift mutations.**

** optional

Step 1:

<input type="checkbox"/> SNV (single nucleotide variation) <input type="checkbox"/> Deletion <input type="checkbox"/> Insertion <input type="checkbox"/> Indel (insertion/deletion) <input checked="" type="checkbox"/> Duplication	Choose the type of variant: duplication
---	---

Step 2:

<input checked="" type="checkbox"/> Reference sequence <input type="checkbox"/> NCBI36 (hg18) <input checked="" type="checkbox"/> GRCh37 (hg19) <input type="checkbox"/> GRCh38 (hg38) <input type="checkbox"/> NM_000546.5 (RNA)	Choose the reference sequence (hg18, hg19, hg38) for genomic entry or the full-length RNA NM_000546.5 for cDNA entry.
---	---

A full description of the TP53 gene and the various TP53 transcripts can be found at the Ensembl website:

(http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=LRG_321;r=17:7668402-7687550)

or at the LRG website

(http://ftp.ebi.ac.uk/pub/databases/lrgex/LRG_321.xml)

Steps 3 and 4:

GRCh37 (hg19)	Duplication
Start position (larger than end position)	End position
7579471	7579471
Wild type allele (coding strand)	Mutant allele (coding strand)
C	e.g., ATGC...

Step3

Choose the start position (first base to be duplicated).

The wild-type allele will be displayed automatically.

Any value outside the range of the various references will generate an error message.

The end position will be set automatically

Step 4 Choose the End position (last base to be duplicated).

all the other steps are performed by Seshat.

Step 5: (Optional)

Most mutations are described using the positive strand of the gene (set by default). If the mutation is described in the negative strand, change the value here.

Step 7 (Optional)

An ID can be given to the analysis.

Step 8

Start the analysis.

9	Button start once all the fields have been filled.
----------	--

Output screen duplication

1	Input data	Summary of the user data.
2	Nomenclature	Official nomenclature of the variant using the three references of the human genome.
3	Description	Important features of the variant. Each item is fully described in the annex document.
4	Full description according to LRG	Full description of the variant in each TP53 transcript according to LRG.
5		Full description of the variant for each TP53 protein isoform according to LRG.
6	TP53 proteins	Description of the variant for the major TP53 protein isoforms; localization of the variant in the various domains of the protein.
7	Comments	Specific comments regarding the TP53 variant. Each item is fully described in the annex document.
10	Summary export	Generate a pdf file with an extensive description of the TP53 variants.
11	Export tables	Generate two TSV files with a full analysis of the variant: see accompanying documents for more information.
12	Back to homepage	Self-explanatory.

Q&A

What is the goal of Seshat?

Seshat performs the following tasks:

- Quality check mutation nomenclature.

- Generates a full description of each variant formatted according to hgvs.

- Generates publication-ready tables.

- Assesses the pathogenicity of each variant according to either general prediction algorithms (Provean, Sift, Polyphen2, FATHMM, MutationAssessor and 7 other algorithms) or algorithms developed exclusively for TP53.

- Displays functional and structural data for each TP53 variant.

What is the extent of Seshat analysis? Is it restricted to the 11 "classical" exons?

Seshat can handle all mutations localized in the TP53 gene, including introns, exons and alternative exons using HG18, HG19 and HG38 nomenclature.

	Start	End
HG18	7536593	7503822
HG19	7595868	7563097
HG38	7692550	7659779

The genomic nomenclature and boundary of TP53 can be found at the LRG website:

http://ftp.ebi.ac.uk/pub/databases/lrgex/LRG_321.xml

The reference sequence is located at this NCBI address and includes 32,772 nucleotides:

http://www.ncbi.nlm.nih.gov/nuccore/NG_017013.2

What information is included in the Seshat TP53 database?

The Seshat TP53 database is different from the TP53 mutation database available at the TP53 website. We have added a lot of additional information to improve the analysis of TP53 mutations.

Among other things, the Seshat database includes:

- Natural SNP data extracted from various databases, such as dbsnp, ExAc or NHLBI (more than 1,000 variants).

- Prediction data for **all** TP53 single nucleotide substitutions.

- Functional data, such as apoptosis, growth arrest or DNA binding (among others) for many TP53 variants.

What type of data is necessary to use Seshat?

Seshat works with CSV files and VCF or MAF files generated by NGS. Input data using protein, cDNA or genomic data can be used. Look at the read me file for more info.

I have a few synonymous variants in the patients that I have analyzed

This is an important issue. Several database have removed all synonymous variants (sSNV) from their data. This is a serious error, as it is now well known that many sSNV are pathogenic with possible defects in RNA splicing, RNA stability or protein folding.

See the following publication for more info:

Soussi T, Taschner PE, Samuels Y (2017) Synonymous Somatic Variants in Human Cancer Are Not Infamous: A Plea for Full Disclosure in Databases and Publications. *Hum Mutat*, **38**: 339–342

Seshat handles and analyzes both sSNV and nSNV and provides a full report of their potential pathogenicity including functional data for many well-known TP53 sSNV.

I have identified a novel TP53 variant not included in the database

This is an important issue, particularly if this variant is a germline variant.

It is a frameshift variant (germline or somatic).

If the mutational event has been verified*, finding a novel deletion or insertion is not infrequent and can be considered to be a pathogenic mutation.

It is a missense variant (germline or somatic).

Analysis of the database shows that the discovery of novel missense variants has decreased considerably over recent years, as most deleterious missense TP53 variants have already been identified. The discovery of a novel missense variant can therefore raise a number of questions, particularly in the case of a germline variant.

Germline Variant: Peripheral blood Lymphocytes

Sequencing larger numbers of individuals will lead to the identification of excessively rare, novel non-pathogenic SNP. Until the pathogenicity of the variant has been clearly demonstrated (segregation with the disease or experimental analysis), these SNP should be considered to be Variants of Unknown Significance.

Various tools can be used to predict the pathogenicity of the mutation, but they have poor sensitivity and specificity for TP53 and must be used with caution. The effect on RNA translation or splicing is not included in predictive software.

You can contact us for further discussion on this variant: p53@free.fr

*Somatic variants**Frozen tissues*

In the absence of any functional information, there is no way to define whether this variant is a driver or passenger mutation.

*Somatic variants**Paraffin-embedded tissues*

DNA sequencing from paraffin-embedded tissues is known to be associated with a high number of artifactual single-nucleotide changes (C:G>T:A). Careful control is necessary to validate this variant.

In the absence of any functional information, there is no way to define whether this variant is a driver or passenger mutation.

* Assuming that sequencing of the genetic material has been carefully performed and controlled.

** Assuming that the somatic origin of this variant has been validated by sequencing normal DNA from the same individual.

Prediction of TP53 mutation pathogenicity: how accurate is this prediction and can it be used in clinical practice?

The prediction of pathogenicity of TP53 variants provided by Seshat is one of the most accurate predictive tools currently available for TP53 variants.

In contrast with all other TP53 mutation databases, including the Cosmic or IARC databases, the Seshat TP53 mutation database has been highly curated to remove all artifactual data linked to sequencing errors.

see Edlund et al. for more info:

Edlund K, Larsson O, Ameer A, Bunikis I, Gyllensten U, Leroy B, Sundstrom M, Micke P, Botling J, 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**: 9551–9556

Pathogenicity has been predicted using specific algorithms developed exclusively for TP53 variants taking three types of non-redundant parameters into account:

Database parameters such as frequency in the database, association with cell lines or in germline among others. These parameters are very accurate and irrefutable, as they are based on more than 70,000 observations .

Exploratory parameters based on multiple experimental data on TP53 mutant loss of function based on the analysis of more than 500 publications.

Predictive parameters based on the use of multiple predictive algorithms.

All popular prediction software use similar algorithms for each protein and do not take into account the specificity of the protein and/or the disease. This leads to loss of specificity, which is unacceptable for clinical applications.

Prediction of loss of function is not synonymous to prediction of pathogenicity. This is an important issue as both prediction are often confused.

TP53 is a multifunctional protein that plays an important role in cancer protection, but which has different properties some of which are not linked to tumor suppression. Therefore, some evolutionarily-conserved residues important for TP53 can be altered without giving rise to a TP53 driver mutation, even when the mutation is predicted to be strongly pathogenic by most predictive software.

For example, TP53 codon 23 is highly conserved in all vertebrates and is important for TP53 regulation. Mutations at codon 23 impair TP53 binding to mdm2 that can lead to defects in cellular growth and these mutations will therefore be counter-selected.

A close examination of TP53 data shows that several highly conserved residues (including codon 23) are totally absent from the database. Unfortunately, such cold spot mutations are always predicted to be pathogenic by predictive software.

The use of three types of non-redundant parameters avoids all of these problems.

As a result of these stringent criteria, we are fairly confident that TP53 variants classified as pathogenic correspond to true driver mutations.

I have unpublished mutations that could be useful for the database

Send us a full description of the mutations using the official nomenclature and we will include your data.

I have used your database for my work. How should I cite Seshat in our publication?

Please use the reference to our last publication and include the url of the website:

Soussi T, Leroy B, Taschner PE (2014) Recommendations for analyzing and reporting TP53 gene variants in the high-throughput sequencing era. *Hum Mutat*, **35**: 766–778

doi: 10.1002/humu.22561.

<https://www.ncbi.nlm.nih.gov.proxy.kib.ki.se/pubmed/24729566>

Thank you

I work in an academic lab and would like to use your database in our analytical pipeline for the analysis of TP53 mutations. Can I obtain Seshat data?

Unfortunately, Seshat data are not publicly available.

What is the future of Seshat?

Seshat data will be regularly updated. Keep an eye on the history file.