interactive
biosoftware

alamut batch 1.4 User Manual

The instructions in this document must be strictly and explicitly followed by qualified and properly trained personnel in order to ensure the proper and safe use of the product(s) described herein. All of the contents of this document must be fully read and understood prior to using such product(s).

Contents

À propos

This user manual describes how to install and use Alamut Batch version 1.4.0 (Feb. 2015).

NOTE
Alamut Batch was previously named *Alamut-HT*

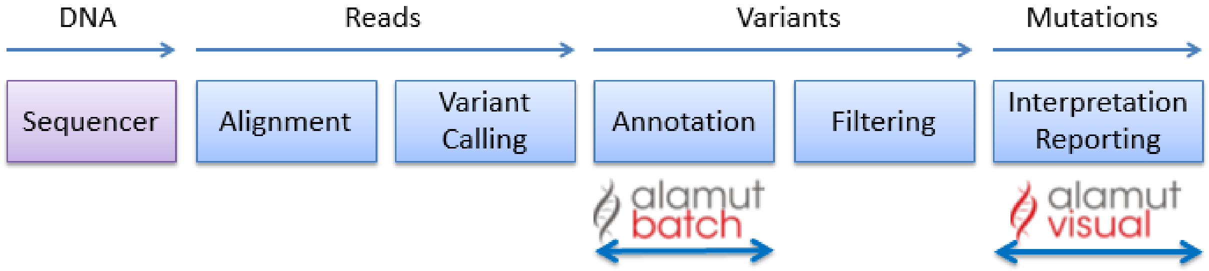
Product description

Alamut Batch is a high-throughput annotation engine for NGS analysis.

Designed for intensive variant analysis workflows, this software enriches raw NGS variants with dozens of annotations including effects on human genes, detailed SNP information, and missense and splicing predictions.

Annotations provided by Alamut Batch are similar to those available in the Alamut® Visual mutation interpretation software. Alamut Batch is able to annotate tens of thousands variants per hour.

This schematic drawing shows where Alamut Batch and Alamut Visual take place in a typical NGS analysis pipeline:



Alamut Batch can be used independently from Alamut Visual. However, results from Alamut Batch can be easily injected into Alamut Visual so as to benefit from its rich feature set, including graphical visualization.

Alamut Batch annotates variants by querying a database storing information about human genes (the Alamut database). Technically, Alamut Batch comes in two versions depending on where the gene database is located:

- The Standalone version uses a locally installed database
- The Client/Server version connects over the internet to our hosted database

Standalone version

The Standalone version of Alamut Batch provides best performance by including in a local installation all software components and the Alamut database required by the annotation process. It is most appropriate for intensive variant annotation needs such those of whole exome analyses.

Alamut Batch Standalone is a Linux command-line program.

Client/Server version

The Client/Server version of Alamut Batch connects remotely to the central Alamut database. Due to internet latency the Client/Server version is slower than the Standalone version but is very easy to install. It is however an efficient solution for moderate variant annotation needs such those of gene panels sequencing analyses.

Alamut Batch Client/Server is available as a command-line program on Windows and Linux operating systems. The software is also available with a GUI frontend on Windows.

System requirements

Standalone version

Alamut Batch Standalone requires the following system specifications:

- 64-bit CentOS 6.4 distribution (or other compatible Linux distribution)
- Python 2.6 or 2.7
- Python MySQLdb package (if access to a local HGMD® Professional database installation is intended)
- OpenSSL client libraries (e.g. RPM package openssl.x86_64)
- 4 GB RAM minimum
- 5 GB hard drive space
- Internet connection required for license control

Client/Server version

Alamut Batch Client/Server requires the following system specifications:

- 64-bit CentOS 6.4 distribution (or other compatible Linux distribution)
- 4 GB RAM minimum
- 100 MB hard drive space
- Internet connection required

Or:

- Windows XP, 7, or 8 (32-bit or 64-bit)
- 2 GB RAM minimum
- 50 MB hard drive space
- Internet connection required

What is new in version 1.4?

Version 1.4 adds annotations from **ClinVar**, **COSMIC** and **ExAC**.

Note that, since a given genomic variant can match multiple ClinVar or COSMIC records, annotations from these datasets are output as lists where each item is separated by a '|' character. Lists in each field are ordered by dataset entries.

For example variant MLH1 NM_000249.2:c.793C>T has 3 entries in ClinVar, yielding the following ClinVar annotation fields:

clinVarIds	RCV000022502.22|RCV000075872.1|RCV000034802.1
clinVarOrigins	germline|germline|germline
clinVarMethods	literature only|research|research
clinVarClinSignifs	Pathogenic|Pathogenic|VUS
clinVarReviewStatus	1|3|1
clinVarPhenotypes	Lynch syndrome ii|Lynch syndrome|Not provided

What was new in version 1.3?

Version 1.3 adds support for the **GRCh38** (hg38) human genome assembly, and includes **1000 genomes** Phase 3 version 5 variant frequencies for five sub-populations (African, East Asian, South Asian, European, American).

What was new in version 1.2?

Version 1.2 introduces the following new features:

- Support for non-protein coding genes now available in the Alamut gene database
- Output annotation lines now include the original variant position provided in the input variant file. (This helps in reconciling variants between the output annotation file and other variant files, which could previously show problematic in case of variant position changes due to application of HGVS rules.)
- VCF quality, filter, information, and genotype fields are now reported in the output even for not-annotated variants
- Annotation can now be restricted to a list of preferred transcripts specified in a gene/transcripts file (`--translist` option)
- Annotation can also be restricted to a range of variants of the input file [`--from` and `--to` options] (not available in the Windows GUI)

Two other new features are specific to the Standalone version:

- Multi-process support: Annotation jobs can now be split among multiple processes on the same computer (`--processes` option)
- Access to local HGMD® Professional database installations has been changed since BIOBASE no longer provides a query API (see Using a local HGMD® Professional database installation)

What was new in version 1.1?

Here are the new features introduced in version 1.1:

- Integration of HGMD (the Human Gene Mutation Database) data, available to HGMD® Professional subscribers
- Integration of NHLBI GO Exome Sequencing Project (ESP) data
- Unannotated variants are now reported in the annotation output file (and in the failed variants output file as well) unless the `--outputannonly` option is specified
- If the new option `--ssIntronicRange <n>` is used, intronic variants located within the specified range `<n>` from the nearest splice site are annotated as 'splice site' in the varLocation annotation field
- Variants can now be filtered by regions of interest defined in a BED format file (`--roilist <ROI list BED file name>`)
- External annotations supplied in variant annotation files can now be integrated in the output (`--extAnnFile <external annotation file name>`)
- Version 1.1.3 adds three output fields reporting validation details of dbSNP entries
- Version 1.1.3 also adds three output fields reporting frequencies of ESP alternate alleles (alternate alleles not always being minor alleles)
- Version 1.1.4 adds an option to allow processing even if the input file has invalid entries
- Version 1.1.5 fixes a bug where variants affecting multiple genes where not processed on all genes
- Version 1.1.6 adds the HGMD variant sub-category output field and fixes a network proxy bug for HTTPS
- Version 1.1.7 brings improvements to the GUI version: all command-line options are now also available in the graphical interface
- With version 1.1.7 it is now possible to input variant alleles that are the same as the transcript allele (e.g. when the genome reference sequence has the minor allele of a SNP and the transcript has the major allele)
- Version 1.1.7 can query a local HGMD database installation
- Version 1.1.8 fixes a bug occurring when a gene cannot be loaded
- Version 1.1.9 features performance improvements and support for mitochondrial variants
- Version 1.1.10 fixes a bug causing software crashes on transcripts where the STOP codon is isolated in a 3’UTR exon
- Version 1.1.11 supports a wider range of VCF variant descriptions (i.e. descriptions that don’t strictly comply with the format specification) and can now output VCF genotype fields of all input samples

Installation

Download the software from http://downloads.interactive-biosoftware.com

The downloaded file is a self-extractable archive on Windows and a tarball on Linux. Extract the contents.

Client/Server GUI frontend (Windows only)

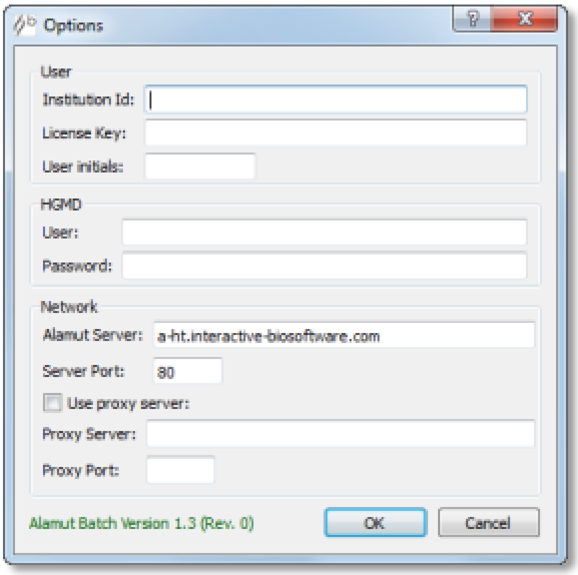
Launch the program Alamut-Batch-UI.exe

Open the Option panel and supply:

- Your Institution ID in the 'Institution' field
- Your license key in the 'Licence Key' field
- User initials as appropriate in the 'User initials' field

If your internet access is behind a proxy, you will also need to supply appropriate proxy settings.

NOTE
The Alamut Server name is 'a-ht.interactive-biosoftware.com' by default. If you are based in North America, please change the server name to 'a-ht-na.interactive-biosoftware.com'.


Options

User
Institution Id:
License Key:
User initials:

HGMD
User:
Password:

Network
Alamut Server: a-ht.interactive-biosoftware.com
Server Port: 80
☐ Use proxy server:
Proxy Server:
Proxy Port:

Alamut Batch Version 1.3 (Rev. 0)

Client/Server command-line program (Windows and Linux)

Edit the `alamut-batch.ini` file and supply:

- Your Institution ID in the 'Institution' field
- Your license key in the 'Licence Key' field
- User initials as appropriate in the 'User' field

NOTE
The Alamut Server name (in field [Network] IBS\Server) is 'a-ht.interactive-

biosoftware.com' by default. If you are based in North America, please change the server name to 'a-ht-na.interactive-biosoftware.com'.

Standalone command-line program (Linux only)

See Installing Alamut Batch Standalone at the end of this document.

Variant Input file

The software takes on input a list of genomic variations, and outputs a list of annotations for each variant, when it is located on a gene available in the Alamut database.

Alamut Batch supports VCF files and tab-delimited files on input.

VCF files — This is the most common format for variant description. Alamut Batch supports VCF v4.0 and later. Note that variants are implicitly processed on the forward strand and that monomorphic references (i.e. entries with no alternate alleles) are not supported.

Tab-delimited files — A specific tab-delimited text format can also be used for variant input. In this format each line should contain the following fields separated by tab characters:

1. Variant id (anything)
2. Chromosome (1-22, X, Y)
3. Genomic position
4. Reference nucleotide(s) (ACGT, or '-' for insertions)
5. Mutated nucleotide(s) (ACGT, or '-' for deletions)
6. Optional strand (1/+ or -1/-), used if `--strand` parameter is set to 0[1]
7. Optional transcript id, used if `--spectrans` parameter is specified
8. Optional user-defined fields (e.g. heterozygosity, number of reads, etc). These fields are not processed but merely reported as-is in the output file.

Empty lines and lines starting with a '#' character are ignored.

Example:

`id00011`	`1`	`23456`	`T`	`A`	`42%`	*T>A substitution*
`id00022`	`9`	`876543`	`-`	`TGA`	`84%`	*TGA insertion*
`id00032`	`5`	`613720`	`AC`	`-`	`2%`	*AC deletion*

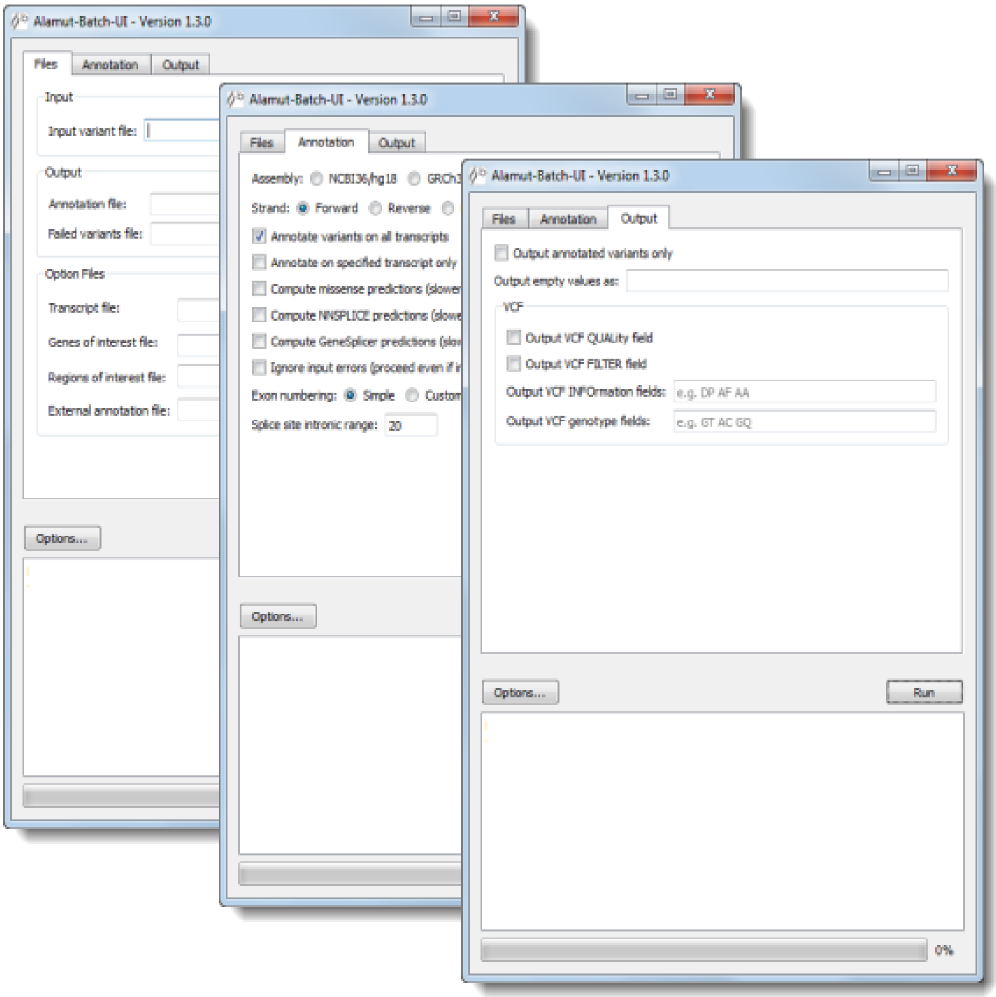
[1] Strand is related to the variant itself, not to the transcript orientation.

Using Alamut Batch

GUI frontend (Windows only)

Launch the program: Alamut-Batch-UI.exe

Program options are spread over three different tabs:



Input

Input variant file:

Output

Annotation file:

Failed variants file:

Option Files

Transcript file:

Genes of interest file:

Regions of interest file:

External annotation file:

Assembly: ○ NCBI36/hg18 ○ GRCh3

Strand: ● Forward ○ Reverse ○

☑ Annotate variants on all transcripts

□ Annotate on specified transcript only

□ Compute missense predictions (slower

□ Compute NNSPLICE predictions (slowe

□ Compute GeneSplicer predictions (slov

□ Ignore input errors (proceed even if i

Exon numbering: ● Simple ○ Custom

Splice site intronic range: 20

□ Output annotated variants only

Output empty values as:

VCF

□ Output VCF QUALity field

□ Output VCF FILTER field

Output VCF INFOrmation fields: e.g. DP AF AA

Output VCF genotype fields: e.g. GT AC GQ

Options are described in section Software Parameters below.

Command-line program (Windows and Linux)

Synopsis:

```
alamut-batch
    [--help]
    [--listgenes <output file name> NCBI36|GRCh37|GRCh38]
     --in <variant file name>
     --ann <annotation file name>
     --unann <unannotated log file name>
1.2  [--from <n>] (start annotating from the nth variant)
1.2  [--to <n>] (annotate up to the nth variant)
     [--assbly NCBI36|GRCh37|GRCh38] (default: GRCh37)
     [--strand 1|-1|0] (default: 1; 0: per variant – not applicable to VCF
                       input)
     [--alltrans] (annotate variants on all transcripts)
     [--spectrans] (annotate variants only on specified per-variant
                   Transcripts – not applicable to VCF input)
1.2  [--translist <transcript file name>] (annotate variants only on listed
                                           preferred transcripts)
     [--glist <gene list file name>] (list of genes of interest)
     [--roilist <ROI list BED file name>] (list of regions of interest)
     [--nomispred] (no missense predictions; faster)
     [--nonnsplice] (no NNSPLICE predictions; faster)
     [--nogenesplicer] (no GeneSplicer predictions; faster)
     [--ignoreInputErrors] (proceed even if input has incorrect entries)
     [--exonnums simple|custom] (default: simple)
     [--ssIntronicRange <n>] (set varLocation as 'splice site' if variant is
                             intronic and within this range)
     [--extAnnFile <external annotation file name>] (include additional
                                                    annotations from external file)
     [--outputannonly] (output only annotated variants in annotation output)
     [--outputVCFQuality]
     [--outputVCFFilter]
     [--outputVCFInfo ID ... ID]
     [--outputVCFGenotypeData ID ... ID]
     [--outputEmptyValuesAs <value>] (e.g. NULL)
     [--hgmdUser <HGMD Pro user name>]
     [--hgmdPasswd <HGMD Pro password>]
     [--proxyserver <proxy server name>]
     [--proxyport <proxy server port number>]
     [--proxyuser <proxy user login>]
     [--proxypasswd <proxy password>]
     [--processes <#processes>] (Standalone version only)
```

(Options flagged as "1.2" were new in version 1.2)

Using the `--listgenes` option puts the program in a special mode making it ouput the unsorted list of genes available in the Alamut database for the given genome assembly.

Options are described in section Software Parameters below.

Software parameters

Input/Output files	Comment	Command line
Variant file	Variant input file full path name (refer to Section "Input file" for details of the file format).	`--in <variant file name>`
Annotation file	Annotation output file full path name (refer to Section "Output file" for details of the file format).	`--ann <annotation file name>`
Failed variants file	Output log file name. This file lists the variants that could not be annotated.	`--unann <unannotated log file name>`
Annotation parameters	**Comment**	**Command line**
Range	Not available in the Windows GUI	`--from <n>` (start annotating from the nth variant) `--to <n>` (annotate up to the nth variant)
Assembly	NCBI36/hg18 or GRCh37/hg19 (The NCBI36/hg18 genome assembly is still supported, but you are strongly encouraged to provide the software with GRCh37/hg19 variations).	`--assbly NCBI36|GRCh37` (default: GRCh37)
Strand	(Not applicable to VCF input) Variants' strand must be explicitly specified, either for the entire input file or on a per variant basis (as specified in column 6 of input file).	`--strand 1|-1|0` 1: forward strand -1: reverse strand 0: per variant (default: 1)
Annotate variants on all transcripts	Each variant will be annotated on all available transcripts if this option is specified. Otherwise only the longest transcript is used.	`--alltrans`
Annotate on specified transcript only	(Not applicable to VCF input) Each variant will be annotated on the transcript specified on a per variant basis (as specified in column 7 of input file).	`--spectrans`
Annotate variants on preferred transcripts listed in specified file	File format described below.	`--translist`
Compute missense predictions	Perform Align GVGD, MAPP and SIFT predictions.	`--nomispred` (cancels default behavior)
Compute NNSPLICE predictions	Perform NNSPLICE predictions.	`--nonnsplice` (cancels default behavior)

Compute GeneSplicer predictions	Perform GeneSplicer predictions.	`--nogenesplicer` (cancels default behavior)
Ignore input errors	Proceed even if input has invalid entries.	`--ignoreInputErrors`
Exon numbering	Simple (sequential) or custom (if available) exon numbering.	`--exonnums simple|custom` (default: simple)
Splice site intronic range	Intronic variants located within the specified range <n> from the nearest splice site are annotated as 'splice site' in the varLocation annotation field	`--ssIntronicRange <n>`
Genes of interest file	List of genes of interest. A file of HGNC gene symbols (1 per line). If this is specified, only variants mapped to the listed genes are annotated.	`--glist <gene list file name>`
Regions of interest file	List of regions of interest (ROIs). A tabulated file where ROIs are described as <chromosome, start, end> (BED format). Only variants located in ROIs are annotated.	`--roilist <ROI list BED file name>`
External annotation file	List of external variant annotations to be reported in output (format described below).	`--extAnnFile <external annotation file name>`
Output parameters	**Comment**	**Command line**
Output annotated variants only	By default variants that cannot be annotated are now (v. 1.1) also reported in the annotation output file. This option cancels this behavior.	`--outputannonly`
VCF quality score	Output VCF QUAL field (applies to VCF input files only)	`--outputVCFQuality`
VCF filter	Output VCF FILTER field (applies to VCF input files only)	`--outputVCFFilter`
VCF information	Output VCF INFO fields specified by a list of IDs, e.g. 'DP AF AA' (applies to VCF input files only)	`--outputVCFInfo ID...ID`
VCF genotype data	Output VCF genotype fields specified by a list of IDs, e.g. 'GT AC GQ' (applies to VCF input files only)	`--outputVCFGenotypeData ID...ID`
Empty values	Empty output fields are populated with specified value, e.g. 'NULL'	`--outputEmptyValuesAs <value>`
HGMD parameters	**Comment**	**Command line**

HGMD® Professional login		--hgmdUser <HGMD Pro user name> --hgmdPasswd <HGMD Pro password>
Proxy parameters	Comment	Command line
Internet proxy options		--proxyserver <proxy server name> --proxyport <proxy server port number> --proxyuser <proxy user login> --proxypasswd <proxy password>

Transcript file format

The input file for preferred transcripts is tab-delimited and requires at least two columns: gene name and transcript name. Multiple transcripts per gene can be specified in additional columns, as in the following example:

```
BRCA1  ->  NM_007294.3
MLH1   ->  NM_000249.3->  NM_001167618.1
```

External annotation files

External variant annotations (e.g. variant pathogenicity status as previously established in the lab) can be integrated in the annotation output.
Variants are described using the chromosome name and genomic-level nomenclature.

Variants and annotations should be supplied in tab-delimited text files using the following format:

- First line: Tab-separated list of annotation labels (preceded by 'chrom' and 'gNomen' for clarity). For example:
  ```
  chrom  ->  gNomen  ->  Class  ->  Freq
  ```
  (where '->' denotes tabulation characters, and 'Class' and 'Freq' are annotation labels)
- Other lines: Tab-separated variant description and annotation values, in the same order as specified in line 1. For example:
  ```
  chr1   ->  g.45800167G>A  ->  Likely pathogenic  ->  0.001
  chr13  ->  g. 32929387T>C ->  Unknown            ->  0.005
  ```

Annotation labels, as supplied in first line, are reported in the first line of the output file. When input variants and externally annotated variants match, the annotation output contains corresponding annotation values.

Note that multiple external variant annotation files can be supplied (using option `--extAnnFile` multiple times).

Using a local HGMD® Professional database installation

If you have a downloaded version of HGMD® Professional you can let Alamut Batch query it locally rather than over the internet. To achieve this you will need to edit the `alamut-batch.ini` file and add an `[HGMD]` section to specify how to connect to the local server, as shown in the following example:

```
[HGMD]
host=192.168.0.1
user=my_hgmd_user
password=my_hgmd_passwd
database=hgmd_pro
```

NOTE
Querying a local HGMD® Pro database is available on Linux only and requires Python and the Python MySQLdb package.

Output

The output of Alamut Batch is a tab-separated file of annotations (1 line per variant or multiple lines per variant if annotation is performed on multiple transcripts). Annotations produced are listed below. User-defined input fields are reported as is in the last output columns.

NOTE
The *Chromosome* field (chrom) was previously the fifth output field. As of v1.2 it now comes as the second field, followed by the new *Variant position* field (pos) that replicates the original variant position given in the input file.

Annotation	Name	Comment
Id	Id	Variant id as supplied in input file
Chromosome	chrom	
Variant position	pos	As supplied in input file
Failed annotation reason	unnnotatedReason	Field not available if option –`outputannonly` is used
Gene symbol	gene	HUGO Gene Nomenclature Committee (HGNC) symbol
Gene id (HGNC)	geneId	HGNC id
Transcript	transcript	e.g.: NM_000249.3
Transcript strand	strand	+/-
Transcript length	transLen	Full cDNA length
Protein	protein	e.g.: NP_000240.1
Uniprot	Uniprot	Uniprot accession, e.g.: P40692
Variant Type	varType	substitution, deletion, insertion, duplication, delins
Variant coding effect	codingEffect	synonymous, missense, nonsense, in-frame, frameshift, start loss, stop loss
Variant location	varLocation	upstream, 5'UTR, exon, intron, 3'UTR, downstream, splice site (see - `--ssIntronicRange` option)
Genome assembly	assembly	
gDNA start	gDNAstart	
gDNA end	gDNAend	
HGVS genomic-level nomenclature	gNomen	e.g.: Chr3(GRCh37):g.37059009A>G
cDNA start	cDNAstart	
cDNA end	cDNAend	
HGVS cDNA-level nomenclature	cNomen	e.g.: NM_000249.3:c.803A>G

HGVS protein-level nomenclature	pNomen	e.g.: p.Glu268Gly
Alt. Protein-level nomenclature	alt_pNomen	Like pNomen except for synonymous variants, e.g.: p.Leu123Leu
Exon	exon	Nearest exon if intronic variant
Intron	intron	
OMIM® id	omimId	

Distance to nearest splice site	distNearestSS	
Nearest splice site type	nearestSSType	5'/3'
WT seq. SpliceSiteFinder score	wtSSFScore	Predictions at nearest splice site
WT seq. MaxEntScan score	wtMaxEntScore	ditto
WT seq. NNSPLICE score	wtNNSScore	ditto
WT seq. GeneSplicer score	wtGSScore	ditto
WT seq. HSF score	wtHSFScore	ditto
Variant seq. SpliceSiteFinder score	varSSFScore	ditto
Variant seq. MaxEntScan score	varMaxEntScore	ditto
Variant seq. NNSPLICE score	varNNSScore	ditto
Variant seq. GeneSplicer score	varGSScore	ditto
Variant seq. HSF score	varHSFScore	ditto
Nearest splice site change	nearestSSChange	Average change predicted by MaxEntScan, NNSPLICE, and HSF
Splicing effect in variation vicinity	localSpliceEffect	New Donor Site, New Acceptor Site, Cryptic Donor Strongly Activated, Cryptic Donor Weakly Activated, Cryptic Acceptor Strongly Activated, Cryptic Acceptor Weakly Activated (*see Section Local splicing effect predictions*)

Protein domain 1	proteinDomain1	
Protein domain 2	proteinDomain2	
Protein domain 3	proteinDomain3	
Protein domain 4	proteinDomain4	

dbSNP variation	rsId	
dbSNP validated variation?	rsValidated	yes/no
dbSNP suspect variation?	rsSuspect	yes/no – Variant flagged as suspect by dbSNP

dbSNP validation labels	rsValidations	e.g.: Cluster/Frequency/1000G
dbSNP number of validation categories	rsValidationNumber	
dbSNP ancestral allele	rsAncestralAllele	
dbSNP variation average heterozygosity	rsHeterozygosity	
dbSNP variation clinical significance	rsClinicalSignificance	
dbSNP variation global Minor Allele Frequency	rsMAF	
dbSNP variation global minor allele	rsMAFAllele	
dbSNP variation sample size	rsMAFCount	

1000 genomes global allele frequency	1000g_AF
1000 genomes allele frequency in African population	1000g_AFR_AF
1000 genomes allele frequency in South Asian population	1000g_SAS_AF
1000 genomes allele frequency in East Asian population	1000g_EAS_AF
1000 genomes allele frequency in European population	1000g_EUR_AF
1000 genomes allele frequency in American population	1000g_AMR_AF

ExAC global allele frequency	exacAllFreq
ExAC allele frequency in African population	exacAFRFreq
ExAC allele frequency in Latino population	exacAMRFreq
ExAC allele frequency in East Asian population	exacEASFreq
ExAC allele frequency in South Asian population	exacSASFreq
ExAC allele frequency in Non-Finnish European population	exacNFEFreq
ExAC allele frequency in Finnish European population	exacFINFreq
ExAC allele frequency in other populations	exacOTHFreq

ExAC homozygosity ratio in African population	exacAFRHmz
ExAC homozygosity ratio in Latino population	exacAMRHmz
ExAC homozygosity ratio in East Asian population	exacEASHmz
ExAC homozygosity ratio in in South Asian population	exacSASHmz
ExAC homozygosity ratio in Non-Finnish European population	exacNFEHmz
ExAC homozygosity ratio in in Finnish European population	exacFINHmz
ExAC homozygosity ratio in other populations	exacOTHHmz
ExAC VCF filter value	exacFilter
ExAC read depth	exacReadDepth

ESP reference allele counts in European American population	espRefEACount
ESP reference allele count in African American population	espRefAACount
ESP reference allele count in all populations	espRefAllCount
ESP alternate allele count in European American population	espAltEACount
ESP alternate allele count in African American population	espAltAACount
ESP alternate allele count in all populations	espAltAllCount
Minor allele frequency in European American population	espEAMAF
Minor allele frequency in African American population	espAAMAF
Minor allele frequency in all populations	espAllMAF
Alternate allele frequency in European American population	espEAAAF
Alternate allele frequency in African American population	espAAAAF
Alternate allele frequency in all populations	espAllAAF

Average sample read depth	espAvgReadDepth	

ClinVar ids	clinVarIds	'|'-separated list
ClinVar origins	clinVarOrigins	'|'-separated list. Possible values: germline, somatic, de novo, maternal, etc
ClinVar methods	clinVarMethods	'|'-separated list. Possible values: clinical testing, research, literature only, etc
ClinVar clinical significances	clinVarClinSignifs	'|'-separated list
ClinVar review status	clinVarReviewStatus	'|'-separated list – Number of stars (0-4)
ClinVar phenotypes	clinVarPhenotypes	'|'-separated list

HGMD mutation id	hgmdId	
HGMD phenotype	hgmdPhenotype	
HGMD PubMed id	hgmdPubMedId	
HGMD sub-category	hgmdSubCategory	DP, DFP, FP, FTV, DM?, DM – see HGMD Documentation website

COSMIC ids	cosmicIds	'|'-separated list
COSMIC tissues	cosmicTissues	'|'-separated list
COSMIC frequencies	cosmicFreqs	'|'-separated list
COSMIC sample counts	cosmicSampleCounts	'|'-separated list

Indels

Inserted nucleotides	insNucs	
Deleted nucleotides	delNucs	

Substitutions

Type	substType	transition, transversion
WT nucleotide	wtNuc	
Variant nucleotide	varNuc	
Nucleotide change	nucChange	
PhastCons score	phastCons	
phyloP	phyloP	

All coding substitutions

WT AA (1 letter)	wtAA_1
WT AA (3 letters)	wtAA_3
WT codon	wtCodon
WT codon frequency	wtCodonFreq
Variant AA (1 letter)	varAA_1
Variant AA (3 letters)	varAA_3
Variant codon	varCodon
Variant codon frequency	varCodonFreq
AA Position	posAA

Missense only

Number of orthologues in alignment	nOrthos
Number of conserved residues in alignment	conservedOrthos
Most distant species in which AA is conserved	conservedDistSpecies

BLOSUM45	BLOSUM45
BLOSUM62	BLOSUM62
BLOSUM80	BLOSUM80
WT AA composition	wtAAcomposition
Variant AA composition	varAAcomposition
WT AA polarity	wtAApolarity
Variant AA polarity	varAApolarity
WT AA volume	wtAAvolume
Variant AA volume	varAAvolume
Grantham distance	granthamDist

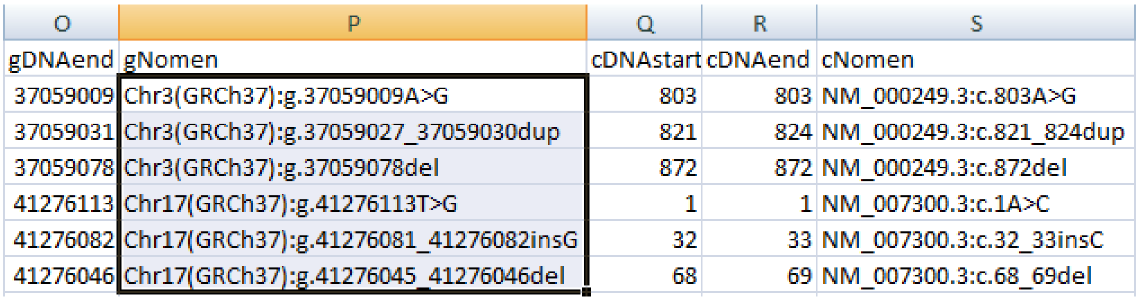
AlignGVGD class	AGVGDclass
AlignGVGD: variation (GV)	AGVGDgv
AlignGVGD: deviation (GD)	AGVGDgd
SIFT prediction	SIFTprediction
SIFT weight	SIFTweight
SIFT median	SIFTmedian

MAPP prediction	MAPPprediction
MAPP p-value	MAPPpValue
MAPP p-value median	MAPPpValueMedian

Viewing annotated variants in Alamut® Visual

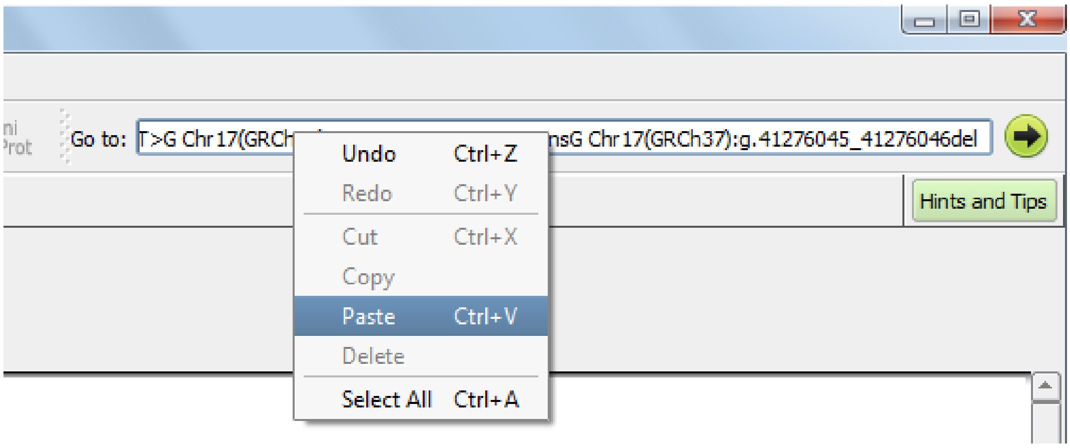
The genomic-level and cDNA-level HGVS descriptions generated by Alamut Batch (annotations gNomen and cNomen) can be easily copied and pasted into Alamut Visual.

Copy a list of HGVS descriptions:

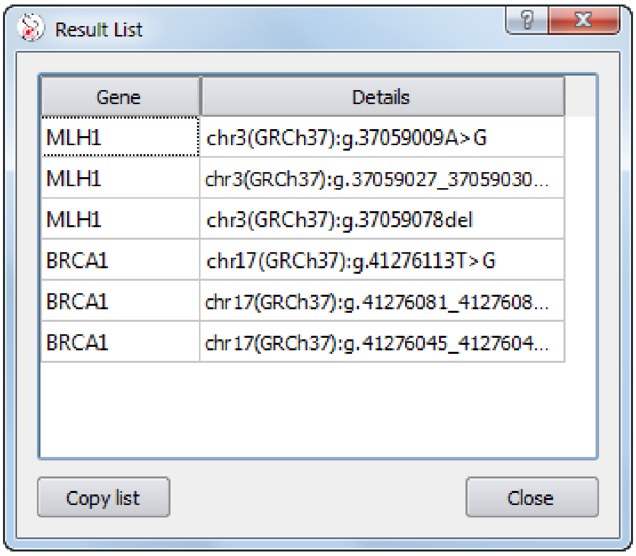


O	P	Q	R	S
gDNAend	gNomen	cDNAstart	cDNAend	cNomen
37059009	Chr3(GRCh37):g.37059009A>G	803	803	NM_000249.3:c.803A>G
37059031	Chr3(GRCh37):g.37059027_37059030dup	821	824	NM_000249.3:c.821_824dup
37059078	Chr3(GRCh37):g.37059078del	872	872	NM_000249.3:c.872del
41276113	Chr17(GRCh37):g.41276113T>G	1	1	NM_007300.3:c.1A>C
41276082	Chr17(GRCh37):g.41276081_41276082insG	32	33	NM_007300.3:c.32_33insC
41276046	Chr17(GRCh37):g.41276045_41276046del	68	69	NM_007300.3:c.68_69del

Then paste it into the Alamut Visual input field:



Variants then show up in a variant list. Double-click on an entry to jump from a variant to another:



Gene	Details
MLH1	chr3(GRCh37):g.37059009A>G
MLH1	chr3(GRCh37):g.37059027_37059030...
MLH1	chr3(GRCh37):g.37059078del
BRCA1	chr17(GRCh37):g.41276113T>G
BRCA1	chr17(GRCh37):g.41276081_4127608...
BRCA1	chr17(GRCh37):g.41276045_4127604...

Local splicing effect predictions

Alamut Batch interprets raw splice site signal recognition by MaxEntScan, NNSPLICE, and Human Splicing Finder (HSF) in the variation vicinity to provide predictions about the creation of new splice sites or the activation of existing cryptic sites.

(Note that this is different from predictions at the nearest splice site, where only raw prediction scores are provided but not interpreted by Alamut Batch.)

This section describes how local splicing effect predictions are computed.

Only the MaxEntScan, NNSPLICE, and HSF splice site predictors are used in the interpretation algorithm. The following thresholds are used to consider or discard raw predictions:

- A MaxEntScan score is deemed significant if > 0
- An NNSPLICE score is deemed significant if > 0.4
- An HSF score is deemed significant if > 60

Let's define a *raw prediction set* as a set of raw predictions at the same position for the same signal. A raw prediction set is deemed significant if at least two of MaxEntScan, NNSPLICE, or HSF predictions are significant.

If, at position *p* (excluding natural splice site positions) there is a significant prediction set both on the wild type sequence and on the mutated sequence, and if the mutated prediction set is significantly higher than the wild type, then Alamut Batch predicts a **cryptic splice site activation**. If the change is less than 3% it is not reported. If it is less than 10% then the activation is reported as **weak**. If it is greater than 10% then it is reported as **strong**.

If, at position *p* there is a significant prediction set on the mutated sequence but not on the wild type sequence, then Alamut Batch predicts a **new splice site creation**.

Installing Alamut Batch Standalone

Alamut Batch Standalone components

Alamut Batch Standalone includes the following components:

1. The Alamut database. It stores all gene-related information used by the software.
2. The `alamut-batch` program. It computes variant annotations based on data provided by the database and results computed by ancillary programs.
3. Ancillary programs. These are external software tools specialized in computing missense and splicing predictions (e.g. SIFT, NNSPLICE).

The Alamut Database

As of version 1.1.11 the Alamut database is supplied as a single compressed file to be used as-is by the `alamut-batch` program (MySQL is no longer required). This file is a snapshot of the live database used by Alamut Visual and the Alamut Batch Client/Server version. Since the live Alamut database is frequently updated, bi-monthly snapshots are provided for Alamut Batch Standalone and can be downloaded from the Alamut website.

The Alamut database includes encrypted gene-related information and must be queried by the `alamut-batch` program only.

The current size of the database is 3.5 GBytes (estimated growth: 3 GBytes/year).

Software Programs

All the required programs are either Linux executables or Python 2.6 scripts. They must all be installed on the same Linux computer.

Ancillary programs include missense and splicing prediction tools that are either provided with the Alamut Batch Standalone package or can be installed separately (see below).

System Requirements

See above.

Installing

Installing Alamut Batch Standalone requires two steps:

- Installing the `alamut_db` database
- Installing software components: Alamut-Batch and ancillary programs

Installing the alamut_db database

Go to the Alamut Batch Standalone section of http://downloads.interactive-biosoftware.com and download the latest database snapshot.

Place the donwload file anywhere in the local filesystem of the computer running Alamut Batch.

Installing Alamut Batch

Go to the Alamut Batch Standalone section of http://downloads.interactive-biosoftware.com and download the latest tarball.

Edit the `alamut-batch.ini` file and supply:

- Your Institution ID in the 'Institution' field
- Your license key in the 'Licence Key' field
- User initials as appropriate in the 'User' field

NOTE
The Alamut Server name (in field [Network] IBS\Server) is 'a-ht.interactive-biosoftware.com' by default. If you are based in North America, please change the server name to 'a-ht-na.interactive-biosoftware.com'.

Set the [Database]/File field to the full path of the downloaded database file.

Installing ancillary programs

All ancillary software programs must be installed in the `alamut-batch-standalone/ancillary` directory:

```
> cd ../alamut-batch-standalone/ancillary
```

SIFT

Download and uncompress:

```
> wget http://sift.jcvi.org/www/sift4.0.3b.tar.gz
> tar zxf sift4.0.3b.tar.gz
```

MAPP (optional)

Download file `MAPP.zip` from http://downloads.interactive-biosoftware.com/?Linux (Section ‘Alamut Batch Standalone’ > ‘Other Downloads’). Unzip this file inside the `ancillary` sub-directory.

NNSPLICE (optional)

Obtain package NNSPLICE0.9 from Martin Reese (mreese@omicia.com) and unpack in the `ancillary` directory.

Note that NNSPLICE requires glibc.i686 (GNU 32-bit libc library).

Other prediction tools

Other tools are either provided with the Alamut Batch distribution (GeneSplicer and MaxEnt) or are embedded inside `alamut-batch` (Align GVGD, SSF, HSF).

Python proxy programs

Two Python proxy programs are needed to ease the communication between Alamut Batch and the ancillary programs: mispred_ht.py and nnsplice_ht.py. Both are provided in the Alamut Batch distribution and must reside in the `ancillary` directory.

The `getHGMD.py` program (also provided in the Alamut Batch distribution) serves as a proxy to connect to a local HGMD® Professional database, if any. This program requires the MySQLdb Python package.

Updating the `alamut_db` database

To update the `alamut_db` database just download the latest snapshot from http://downloads.interactive-biosoftware.com and edit the `alamut-batch.ini` file to change the [Database]/File field appropriately.