

Data dictionary of Identitag

Protein

This table stores information about proteins homologous to the transcript sequences stored in Transcript table

Field	Signification	Type	Description	Foreign key
Id	Unique identification number of the protein	int(8)	NOT NULL PRIMARY KEY AUTO INCREMENT	
Sptrembl_id	Identification number of the protein in Swissprot or TrEMBL databank	varchar(30)	NOT NULL	
Sptrembl_ac	Primary accession number of the protein in Swissprot or TrEMBL databank	varchar(30)	NOT NULL	
Description	Description field (DE) in Swissprot or TrEMBL databank	text		
Organism_species	Organism species field (OS) in Swissprot or TrEMBL databank	text		
Databank	Databank from which this protein originates	varchar(20)	NOT NULL	

Species_Observed_Tag

This table stores information about tags from SAGE libraries (i.e. tags experimentally obtained)

Field	Signification	Type	Description	Foreign key
Id	Unique identification number of the observed tag	int(8)	NOT NULL PRIMARY KEY AUTO INCREMENT	
Library_id	Unique identification number of the library from which this tag was derived	int(8)	NOT NULL	Species_SAGE_Library
Sequence	Tag sequence. Contains the anchoring enzyme restriction site (14 bp total length or 21 bp for long SAGE tags)	varchar(22)	NOT NULL	
Tag_number	Number of time this tag was observed in the corresponding library identified by Library_id	int(4)	NOT NULL	

Species_SAGE_Library

This table stores experimental conditions for SAGE experiment

Field	Signification	Type	Description	Foreign key
Id	Unique identification number of the SAGE library	int(8)	NOT NULL PRIMARY KEY AUTO INCREMENT	
Name	Name of the SAGE library	varchar(20)	NOT NULL	
Library_Submitter	Submitter of the SAGE library (First name Last name)	varchar(50)	NOT NULL	
Cells_Submitter	Submitter of the cells preparation before SAGE experiment (First name Last name)	varchar(50)	NOT NULL	
Experiment_number	Cells preparation experiment number	varchar(50)	NOT NULL	
Strain	Strain from which the cells come from	varchar(10)	NOT NULL	
Age	Age of the cells (in days)	int(3)	NOT NULL	
Treatment	Short explanation of the cells treatment before SAGE experiment	varchar(255)	NOT NULL	
Kit_number	Identification number of the SAGE kit	varchar(30)	NOT NULL	
Enzyme	Anchoring enzyme used for SAGE experiment	varchar(10)	NOT NULL	

Species_Transcript

This table stores information about transcript sequences

Field	Signification	Type	Description	Foreign key
Id	Unique identification number of the transcript	int(8)	NOT NULL PRIMARY KEY AUTO INCREMENT	
Sequence_id	Transcript identification number in the databank from which it was derived	varchar(40)	NOT NULL	
Source	Name of the databank this transcript originates	varchar(15)	NOT NULL	
AATAAA_signal	1 if an AATAAA signal is present in the transcript sequence, else 0	tinyint(1)	NOT NULL	
AATAAA_position	AATAAA signal position in the transcript sequence (in bp from the beginning of the transcript sequence)	int(3)	NOT NULL	
ATTAAA_signal	1 if an ATTAAA signal is present in the transcript sequence, else 0	tinyint(1)	NOT NULL	
ATTAAA_position	ATTAAA signal position in the transcript sequence (in bp from the beginning of the transcript sequence)	int(3)	NOT NULL	
A_number_tail	Length of the longest polyA stretch among the 50 last bp of the transcript sequence	tinyint(4)	NOT NULL	
3_label	1 if there is a 3 prime label at the end of the transcript sequence, else 0	tinyint(1)	NOT NULL	
Sequence_length	Length of the transcript sequence (in bp)	smallint(6)	NOT NULL	
Description	Description field (DE) if the transcript sequence originates a databank containing a such field (e.g. Genbank)	text		
Sequence	Transcript sequence	text	NOT NULL	

Species_Transcript_Protein

This is an association table between a transcript (query sequence) stored in Transcript table and its best BLASTX hits (harboring an Evalue less than 0.001) stored in Protein table (subject sequence). This table stores corresponding BLASTX parameters.

Field	Signification	Type	Description	Foreign key
SpeciesTranscript_id	Unique identification number of the transcript sequence	int(8)	NOT NULL PRIMARY KEY	Species_Transcript
Protein_id	Unique identification number of the protein	int(8)	NOT NULL PRIMARY KEY	Protein
Blastx_identity_percentage	Identity percentage between the two sequences	float		
Blastx_alignment_length	Length of the alignment between the two sequences	int(6)		
Blastx_mismatches	Number of mismatches in the alignment	int(6)		
Blastx_gap_openings	Number of gap openings in the alignment	int(6)		
Blastx_query_start	Position of the alignment beginning in the query sequence (in bp)	int(6)		
Blastx_query_end	Position of the alignment end in the query sequence (in bp)	int(6)		
Blastx_subject_start	Position of the alignment beginning in the subject sequence (in bp)	int(6)		
Blastx_subject_end	Position of the alignment end in the subject sequence (in bp)	int(6)		
Blastx_evalue	BLASTX e-value	float		
Blastx_score	BLASTX score	float		

PRIMARY KEY(SpeciesTranscript_id, Protein_id)

Species_Virtual_Observed_Tag

This is an association table between an experimentally observed tag and a virtual one

Field	Signification	Type	Description	Foreign key
SpeciesObservedTag_id	Unique identification number of the observed tag	int(8)	NOT NULL PRIMARY KEY	Species_Observed_Tag
SpeciesVirtualTag_id	Unique identification number of the virtual tag	int(8)	NOT NULL PRIMARY KEY	Species_Virtual_Tag

PRIMARY KEY(SpeciesObservedTag_id, SpeciesVirtualTag_id)

Species_Virtual_Tag

This table describes virtual tags generated from transcript sequences

Field	Signification	Type	Description	Foreign key
Id	Unique identification number of the virtual tag	int(8)	NOT NULL PRIMARY KEY AUTO INCREMENT	
Sequence_short	Tag sequence obtained by virtual SAGE using BsmF1 as a tagging enzyme. Contains the anchoring enzyme restriction site (14 bp total length)	varchar(15)	NOT NULL	
Sequence_long	Tag sequence obtained by virtual long SAGE using MmeI as a tagging enzyme. Contains the anchoring enzyme restriction site (21 bp total length)	varchar(22)		
Enzyme	Anchoring enzyme used to generate this virtual tag	varchar(10)	NOT NULL	
Enzyme_site_position	last if the tag corresponds to the last recognition site in the transcript sequence, before if the tag corresponds to the before-to-last recognition site in the transcript sequence	enum('last', 'before')	NOT NULL	

Species_Virtual_Tag_Transcript

This is an association table between a transcript sequence and the virtual tags extracted from this sequence

Field	Signification	Type	Description	Foreign key
SpeciesVirtualTag_id	Unique identification number of the virtual tag	int(8)	NOT NULL	
SpeciesTranscript_id	Unique identification number of the transcript from wich this virtual tag was derived	int(8)	NOT NULL	

PRIMARY KEY(SpeciesVirtualTag_id, SpeciesTranscript_id)