The COLLECTF external submission guide (v1.2)

COLLECTF, a play on words using the French *collectif* [collective] and the acronym for transcription factor [TF], is a database of prokaryotic transcription factor binding sites (TFBS). The main aim of COLLECTF is to provide high-quality, manually-curated information on the experimental evidence for transcription factor binding sites, and to map these onto reference bacterial genomes for ease of access and processing.

This document is intended to provide external submitters with a companion guide for the submission process. Additional information is available on the COLLECTF wiki. COLLECTF is accessible at http://collectf.umbc.edu.

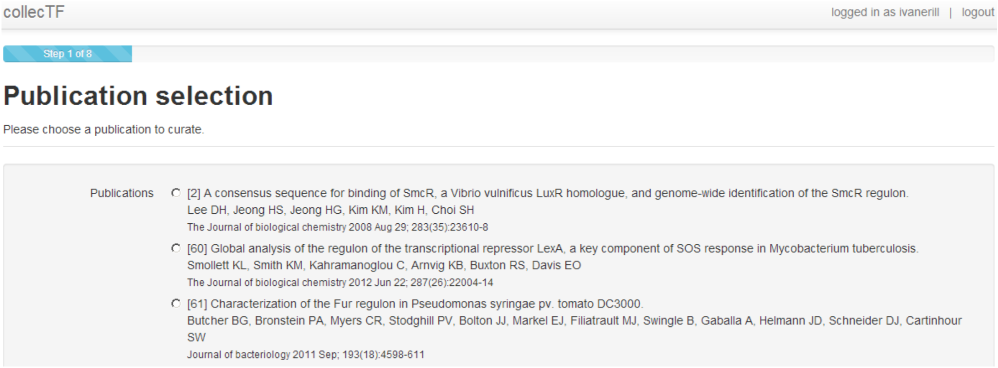
The COLLECTF submission process

Data

COLLECTF uses data from only one type of source: published experimental evidence on transcription factor binding sites. COLLECTF distinguishes between two types of experimental support: evidence of binding (e.g. EMSA) and evidence of TF-mediated regulation (e.g. β-galactosidase assay). Identification of TF-binding sites through in silico means is recorded as part of the curation process, but not admitted as the single source of evidence for a TF-binding site. *Please do not submit data without some form of experimental evidence*.

Step 0: Publication selection

The submission process starts with the submitter selecting a publication for curation. You must submit a publication for curation before starting a curation. You can submit several publications for curation.



collecTF

logged in as ivanerill | logout

Step 1 of 8

Publication selection

Please choose a publication to curate.

Publications

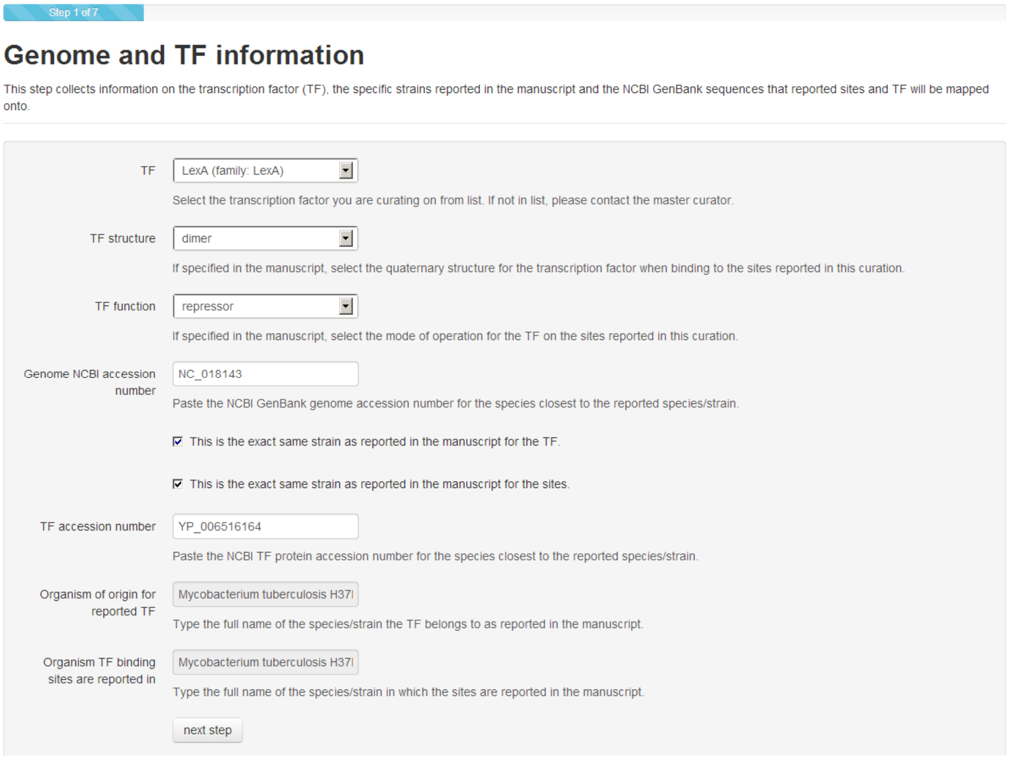
○ [2] A consensus sequence for binding of SmcR, a Vibrio vulnificus LuxR homologue, and genome-wide identification of the SmcR regulon.
Lee DH, Jeong HS, Jeong HG, Kim KM, Kim H, Choi SH
The Journal of biological chemistry 2008 Aug 29; 283(35):23610-8

○ [60] Global analysis of the regulon of the transcriptional repressor LexA, a key component of SOS response in Mycobacterium tuberculosis.
Smollett KL, Smith KM, Kahramanoglou C, Arnvig KB, Buxton RS, Davis EO
The Journal of biological chemistry 2012 Jun 22; 287(26):22004-14

○ [61] Characterization of the Fur regulon in Pseudomonas syringae pv. tomato DC3000.
Butcher BG, Bronstein PA, Myers CR, Stodghill PV, Bolton JJ, Markel EJ, Filiatrault MJ, Swingle B, Gaballa A, Helmann JD, Schneider DJ, Cartinhour SW
Journal of bacteriology 2011 Sep; 193(18):4598-611

Step 1: Genome and TF information

Once a publication has been selected, the submitter must link the reported species (both for the sites and the transcription factor) to sequences present in the NCBI RefSeq database, by providing RefSeq accession number for a genome file (e.g. NC_005363.1; including version number) and the protein corresponding to the TF (e.g. NP_970244; no version number). This is often a simple step, but can get more complex if the sequence for the exact strain used in your work is not available as an NCBI RefSeq record. Please try to identify a parental or related strain among those in NCBI RefSeq genomes. If there is no clear way to identify a surrogate genome in NCBI RefSeq, please use that of a common lab strain (most likely already in the database) for data submission.



Genome and TF information

This step collects information on the transcription factor (TF), the specific strains reported in the manuscript and the NCBI GenBank sequences that reported sites and TF will be mapped onto.

TF: LexA (family: LexA)
Select the transcription factor you are curating on from list. If not in list, please contact the master curator.

TF structure: dimer
If specified in the manuscript, select the quaternary structure for the transcription factor when binding to the sites reported in this curation.

TF function: repressor
If specified in the manuscript, select the mode of operation for the TF on the sites reported in this curation.

Genome NCBI accession number: NC_018143
Paste the NCBI GenBank genome accession number for the species closest to the reported species/strain.

☑ This is the exact same strain as reported in the manuscript for the TF.

☑ This is the exact same strain as reported in the manuscript for the sites.

TF accession number: YP_006516164
Paste the NCBI TF protein accession number for the species closest to the reported species/strain.

Organism of origin for reported TF: Mycobacterium tuberculosis H37l
Type the full name of the species/strain the TF belongs to as reported in the manuscript.

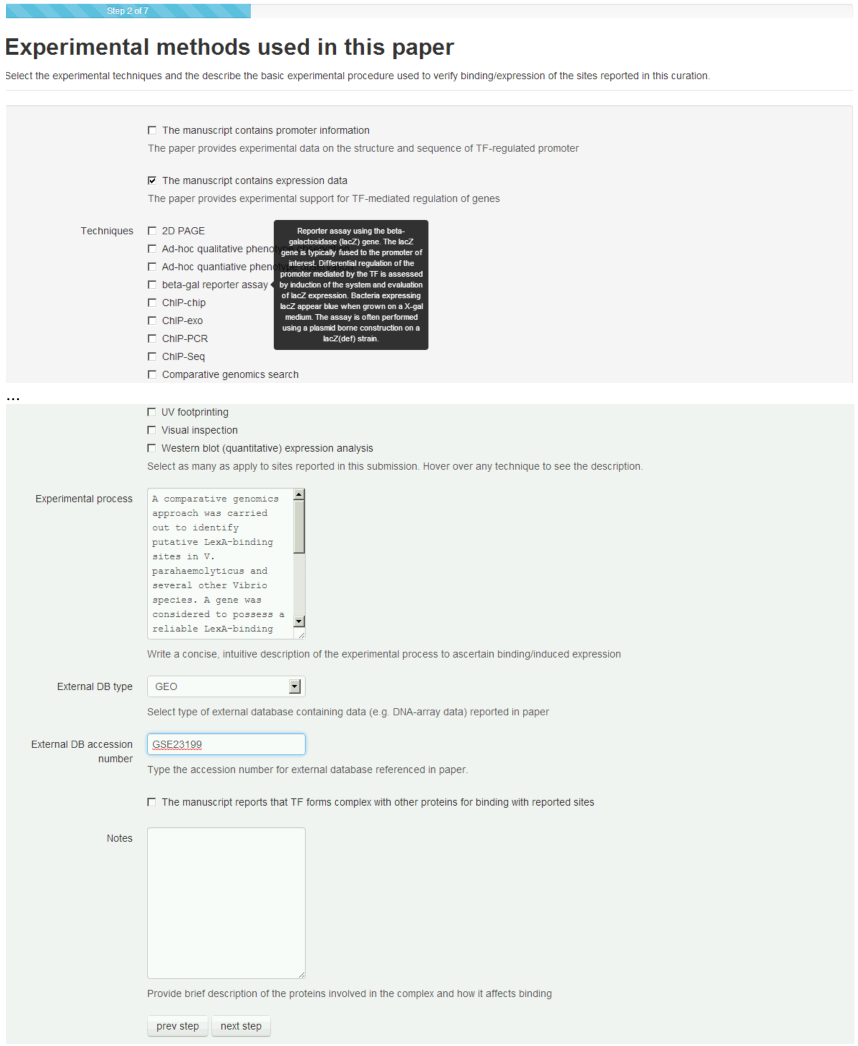
Organism TF binding sites are reported in: Mycobacterium tuberculosis H37l
Type the full name of the species/strain in which the sites are reported in the manuscript.

In this screen you must first select the *TF* you are reporting on. If you are reporting that the TF acts in some multi-meric form, please indicate this too in the `TF structure` drop list. COLLECTF associates activation and repression information globally for each curation. This means that if your work describes both activated and repressed sites, you must submit two curations, selecting the appropriate `TF function` in each case and including, only the activated/repressed sites in each curation. If a site is both repressed and activated, it should be present in both curations. While this may sound tedious, the submission system will pre-populate all fields in your second (or third…) submission, facilitating the process enormously.

If the work you are reporting uses a strain different from the selected RefSeq genome/TF, please type/paste the original strain in the `Organism of origin...` and `Organism TF binding sites...` text fields. This allows us to keep track of the correspondence between reported and mapped strains.

Step 2: Experimental methods

COLLECTF is all about gathering and validating information on experimentally-validated TF-binding sites. The next two steps in the curation process specifically target these two fundamental points. Step 2 requires that you report <u>*all* the techniques used in the paper to verify the TFBS</u> that are being <u>reported *in this submission*</u>. In this step we also ask that you provide a <u>*brief written summary*</u> of the process used to verify the submitted TFBS (not the overall experimental process, but just how the selected experimental techniques were combined to define reported TFBS)[1]. You can also indicate whether the paper contains promoter information (e.g. location of transcriptional start) or expression data (evidence of TF-mediated regulation), and provide external database accession numbers for expression data (e.g. GEO accession numbers), or details on whether the TF forms complex with other molecules in order to bind.



Experimental methods used in this paper

Select the experimental techniques and the describe the basic experimental procedure used to verify binding/expression of the sites reported in this curation.

□ The manuscript contains promoter information

The paper provides experimental data on the structure and sequence of TF-regulated promoter

☑ The manuscript contains expression data

The paper provides experimental support for TF-mediated regulation of genes

Techniques
□ 2D PAGE
□ Ad-hoc qualitative pheno
□ Ad-hoc quantiative pheno
□ beta-gal reporter assay
□ ChIP-chip
□ ChIP-exo
□ ChIP-PCR
□ ChIP-Seq
□ Comparative genomics search

Reporter assay using the beta-galactosidase (lacZ) gene. The lacZ gene is typically fused to the promoter of interest. Differential regulation of the promoter mediated by the TF is assessed by induction of the system and evaluation of lacZ expression. Bacteria expressing lacZ appear blue when grown on a X-gal medium. The assay is often performed using a plasmid borne construction on a lacZ(def) strain.

...

□ UV footprinting
□ Visual inspection
□ Western blot (quantitative) expression analysis

Select as many as apply to sites reported in this submission. Hover over any technique to see the description.

Experimental process: A comparative genomics approach was carried out to identify putative LexA-binding sites in V. parahaemolyticus and several other Vibrio species. A gene was considered to possess a reliable LexA-binding

Write a concise, intuitive description of the experimental process to ascertain binding/induced expression

External DB type: GEO

Select type of external database containing data (e.g. DNA-array data) reported in paper

External DB accession number: GSE23199

Type the accession number for external database referenced in paper.

□ The manuscript reports that TF forms complex with other proteins for binding with reported sites

Notes:

Provide brief description of the proteins involved in the complex and how it affects binding

[1] For instance: *"Sites were first identified using a computer search, then binding was validated with EMSA. TF-mediated expression was confirmed with β-gal assays on w-t vs. tf- mutant"*. See curations in the database for examples.

Step 3: Reported sites

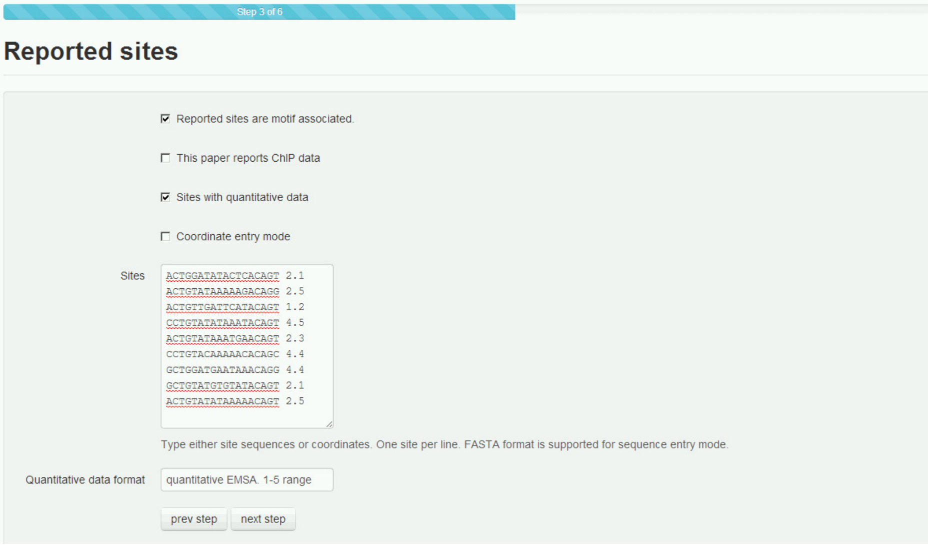
In this step, you will enter the primary information for CollecTF: binding sites reported in your work *using the techniques specified in Step 2*. Notice that if your work describes different groups of sites using different sets of techniques, *you should accordingly create multiple independent submissions*.

Motif-associated

TF-binding sites can be defined at two different levels. By definition, a TF-binding site is simply a (relatively short) stretch of DNA to which a transcription factor is shown to bind. Many TFs target known specific sequence patterns in the DNA that lead to aligned collections of sites (motifs), providing a much more concise definition of TF-binding site. In CollecTF we refer to the latter as motif-associated and the former as non-motif associated. If you are confident that the sites you report conform to a known motif or you demonstrate that they do through experimental work (e.g. site-directed mutagenesis), you should check the `Reported sites are motif associated` checkbox.

Coordinates and quantitative data

Sites can be entered as sequences or using genome coordinates (if you already mapped them to the reference strain in your work), and the appropriate box should be checked. Sites should be entered one per line (FASTA format is also accepted for sequence entry). If you report quantitative data for sites (e.g. peak intensities, estimated K_d), please check the `Sites with quantitative data` checkbox. Quantitative data can then be appended with a tab/space after the sequence/coordinate entry. A brief description of its nature (method used and range of quantitative data) should be entered in the `Quantitative data format` textbox.



Reported sites

☑ Reported sites are motif associated.

☐ This paper reports ChIP data

☑ Sites with quantitative data

☐ Coordinate entry mode

Sites

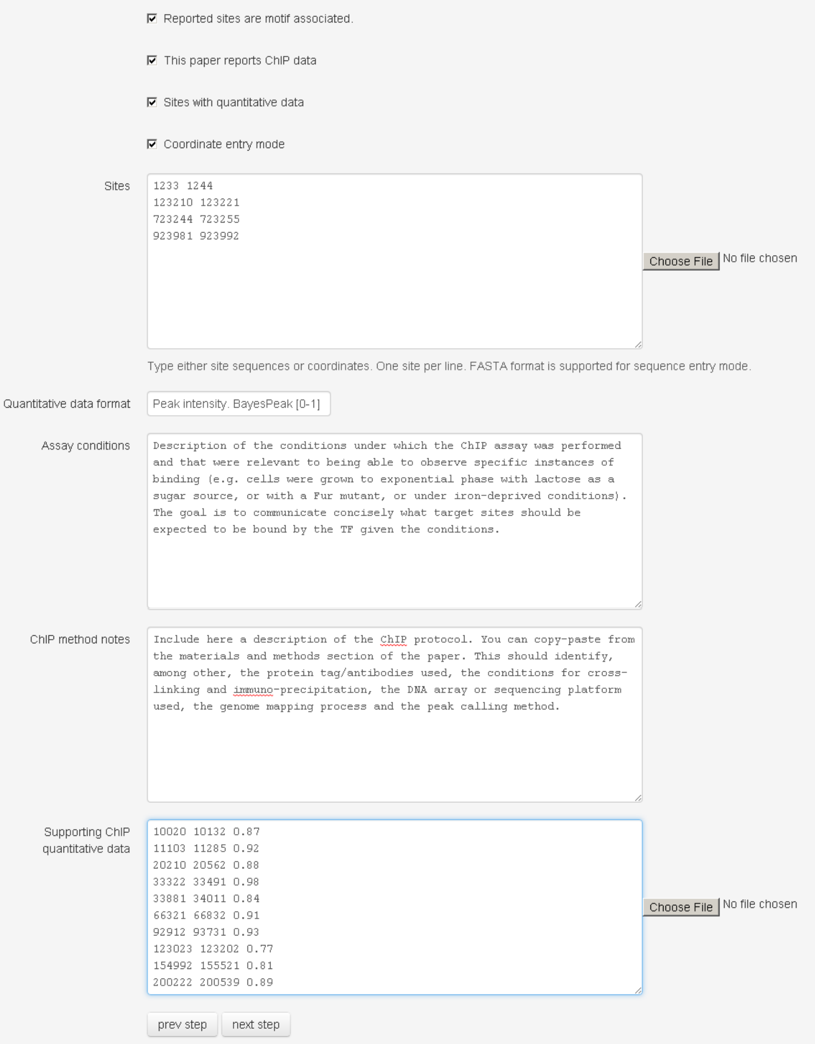
```
ACTGGATATACTCACAGT 2.1
ACTGTATAAAAAGACAGG 2.5
ACTGTTGATTCATACAGT 1.2
CCTGTATATAAATACAGT 4.5
ACTGTATAAATGAACAGT 2.3
CCTGTACAAAAACACAGC 4.4
GCTGGATGAATAAACAGG 4.4
GCTGTATGTGTATACAGT 2.1
ACTGTATATAAAAACAGT 2.5
```

Type either site sequences or coordinates. One site per line. FASTA format is supported for sequence entry mode.

Quantitative data format: quantitative EMSA. 1-5 range

ChIP data

COLLECTF also captures ChIP peak data. When the `This paper reports ChIP data` box is checked, the submission process will request additional data.



☑ Reported sites are motif associated.

☑ This paper reports ChIP data

☑ Sites with quantitative data

☑ Coordinate entry mode

Sites

```
1233 1244
123210 123221
723244 723255
923981 923992
```

Type either site sequences or coordinates. One site per line. FASTA format is supported for sequence entry mode.

Quantitative data format: Peak intensity. BayesPeak [0-1]

Assay conditions: Description of the conditions under which the ChIP assay was performed and that were relevant to being able to observe specific instances of binding (e.g. cells were grown to exponential phase with lactose as a sugar source, or with a Fur mutant, or under iron-deprived conditions). The goal is to communicate concisely what target sites should be expected to be bound by the TF given the conditions.

ChIP method notes: Include here a description of the ChIP protocol. You can copy-paste from the materials and methods section of the paper. This should identify, among other, the protein tag/antibodies used, the conditions for cross-linking and immuno-precipitation, the DNA array or sequencing platform used, the genome mapping process and the peak calling method.

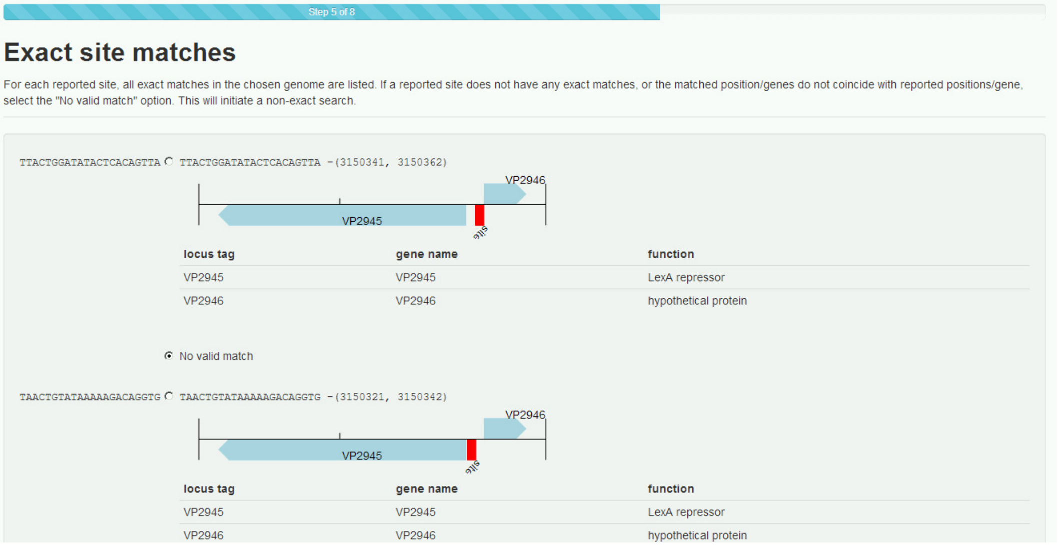
Supporting ChIP quantitative data

```
10020 10132 0.87
11103 11285 0.92
20210 20562 0.88
33322 33491 0.98
33881 34011 0.84
66321 66832 0.91
92912 93731 0.93
123023 123202 0.77
154992 155521 0.81
200222 200539 0.89
```

The `Assay conditions` field should report the specific biological conditions on which the ChIP assay was carried out (e.g. iron deprivation). The `ChIP method notes` field should detail specific ChIP protocols (cross-linking method, sequencing, etc.) as reported in the *Materials and Methods* section of the manuscript. The `Supporting ChIP quantitative data` field accepts peak data with quantitative peak intensities and allows you to assign this quantitative data to reported motif associated sites. If this field is populated, the COLLECTF submission system will scan peak data for instances of reported motif associated sites (in the `Sites` field) and automatically associate the peak value to the motif associated site.

Step 4: Verify sites (exact)

After you enter the sites, the COLLECTF submission system will download the specified genome sequence and search for entered sites. The sites are reported back to the curator specifying their location in the sequence and nearby genes. Gene annotation details can be accessed by hovering over any gene locus. You can use this information to verify that the sites identified by the COLLECTF submission system in the NCBI RefSeq genome sequence correspond to the sites you report in the paper.


Step 5 of 8

Exact site matches

For each reported site, all exact matches in the chosen genome are listed. If a reported site does not have any exact matches, or the matched position/genes do not coincide with reported positions/gene, select the "No valid match" option. This will initiate a non-exact search.

TTACTGGATATACTCACAGTTA ○ TTACTGGATATACTCACAGTTA –(3150341, 3150362)

locus tag	gene name	function
VP2945	VP2945	LexA repressor
VP2946	VP2946	hypothetical protein

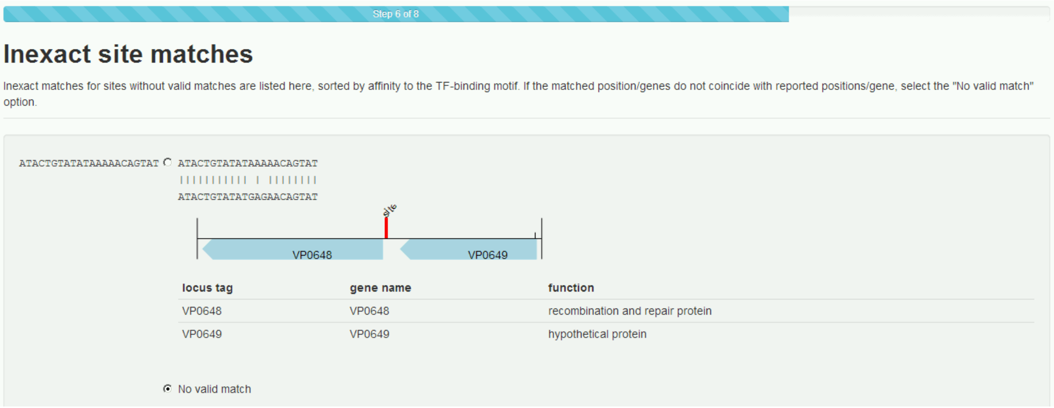
● No valid match

TAACTGTATAAAAAGACAGGTG ○ TAACTGTATAAAAAGACAGGTG –(3150321, 3150342)

locus tag	gene name	function
VP2945	VP2945	LexA repressor
VP2946	VP2946	hypothetical protein

Step 6: Verify sites (inexact)

In some cases, especially if using a sequence that is not an exact match to the reported strain, some sites may not be found using an exact search. In this case, the COLLECTF submission system will use the available evidence to construct a scoring matrix and search the genome for slightly inexact matches (up to two mismatches away from the reported site). These will be reported in the same way as exact matches and you will be asked to validate them in the same manner.


Step 6 of 8

Inexact site matches

Inexact matches for sites without valid matches are listed here, sorted by affinity to the TF-binding motif. If the matched position/genes do not coincide with reported positions/gene, select the "No valid match" option.

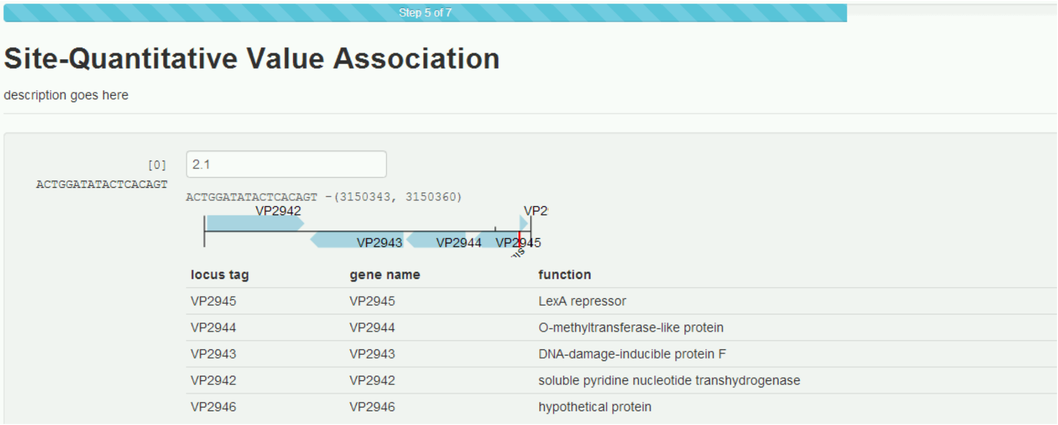
```
ATACTGTATATAAAAACAGTAT ○ ATACTGTATATAAAAACAGTAT
                         ||||||||||| | ||||||||
                         ATACTGTATATGAGAACAGTAT
```

locus tag	gene name	function
VP0648	VP0648	recombination and repair protein
VP0649	VP0649	hypothetical protein

● No valid match

Step 6: Verify quantitative data

If quantitative data has been entered for sites, the COLLECTF system will also display it along with sites in their genomic context for verification and editing.



Step 5 of 7

Site-Quantitative Value Association

description goes here

[0] 2.1

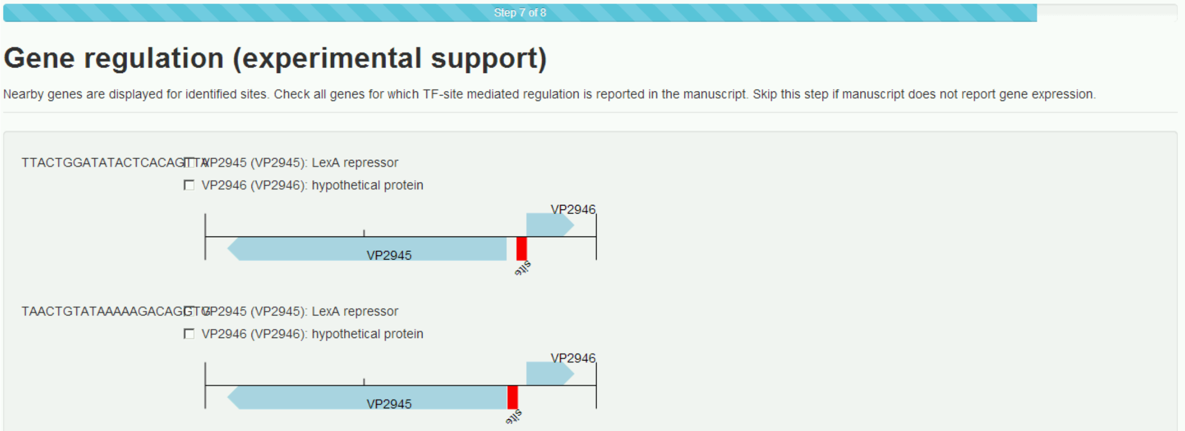
ACTGGATATACTCACAGT

ACTGGATATACTCACAGT -(3150343, 3150360)

locus tag	gene name	function
VP2945	VP2945	LexA repressor
VP2944	VP2944	O-methyltransferase-like protein
VP2943	VP2943	DNA-damage-inducible protein F
VP2942	VP2942	soluble pyridine nucleotide transhydrogenase
VP2946	VP2946	hypothetical protein

Step 7: Gene regulation

If the manuscript reports experimental evidence for TF-mediated regulation of target genes through TFBS, the COLLECTF submission system will ask you to specify, for each reported site, which genes have been shown to be regulated by the TF.



Step 7 of 8

Gene regulation (experimental support)

Nearby genes are displayed for identified sites. Check all genes for which TF-site mediated regulation is reported in the manuscript. Skip this step if manuscript does not report gene expression.

TTACTGGATATACTCACAGTTTA ☐ VP2945 (VP2945): LexA repressor
☐ VP2946 (VP2946): hypothetical protein

TAACTGTATAAAAAGACAGGTG ☐ VP2945 (VP2945): LexA repressor
☐ VP2946 (VP2946): hypothetical protein

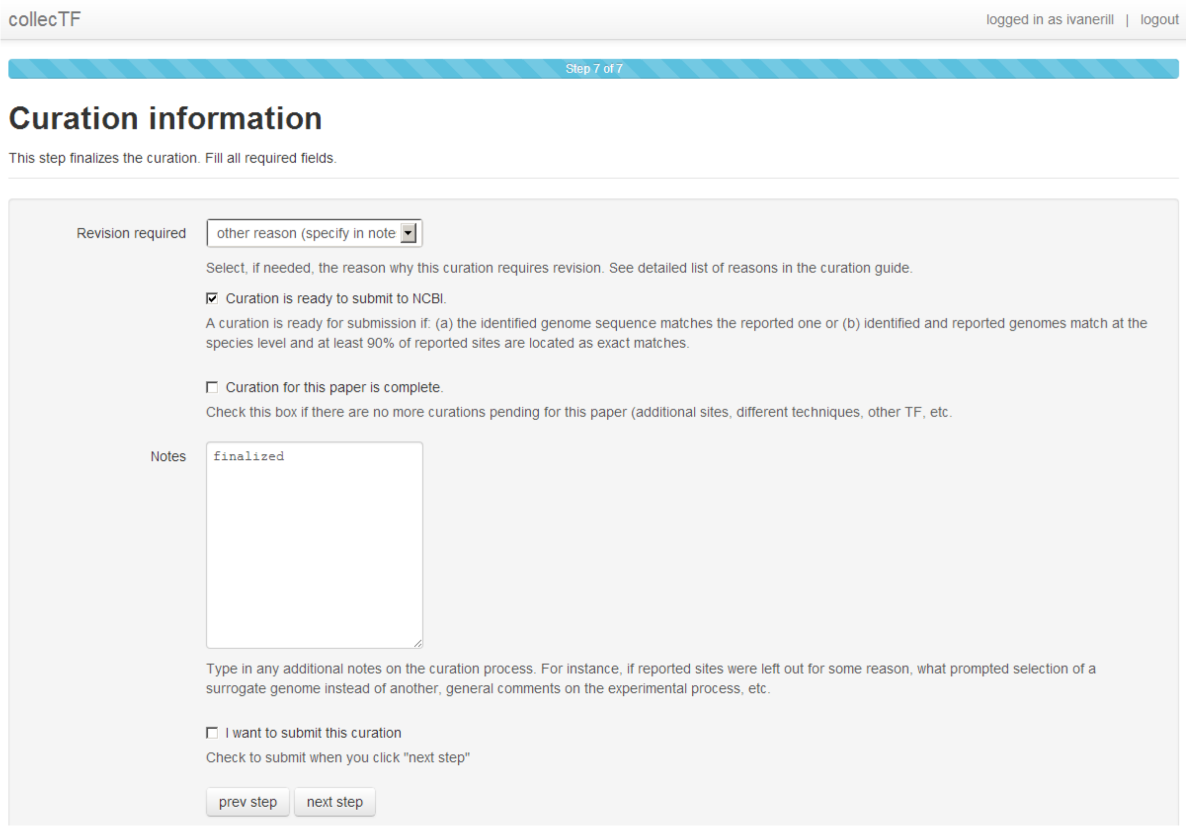
Step 8: Curation information

The submission process ends with a final assessment of the curation. You will be asked whether the submission requires review (`Revision required`). Checking this option is indicated in several circumstances. For instance, it is quite possible that no appropriate sequence has been located in NCBI to perform a valid curation. In this case, the curation is marked for revision. The TFBS data is stored, but it will not be linked to a RefSeq sequence until a matching RefSeq record is posted.

You will also be asked whether the curation is considered valid for submission to NCBI. Curations will only be *considered for submission to NCBI if the sequence for the exact reported strain is available at NCBI or if a sequence matching the species of the reported strain is available and at least 90% of the sites you report have been located in the reference RefSeq record as exact matches*.

The system also requires that you specify if the `Curation for this paper is complete`. Do not check this box if, for instance, you must report *additional sites with different support* that need to be reported in a subsequent curation, you must report *additional sites in a different chromosome or for a different TF* or if you need to report *additional sites under a different mode of regulation* (e.g. repressed instead of activated). All these situations require multiple curations. The COLLECTF submission system will pre-populate fields to facilitate this process.

Finally, after you check `I want to submit this curation` and click `Next`, the system will ask that you verify the submission (a new window summarizing the submission will pop up for inspection).



Curation information

This step finalizes the curation. Fill all required fields.

Revision required: other reason (specify in note

Select, if needed, the reason why this curation requires revision. See detailed list of reasons in the curation guide.

☑ Curation is ready to submit to NCBI.

A curation is ready for submission if: (a) the identified genome sequence matches the reported one or (b) identified and reported genomes match at the species level and at least 90% of reported sites are located as exact matches.

□ Curation for this paper is complete.

Check this box if there are no more curations pending for this paper (additional sites, different techniques, other TF, etc.

Notes: finalized

Type in any additional notes on the curation process. For instance, if reported sites were left out for some reason, what prompted selection of a surrogate genome instead of another, general comments on the experimental process, etc.

□ I want to submit this curation

Check to submit when you click "next step"

Once a submission is completed, the data is uploaded to COLLECTF. The submission will be then reviewed by a COLLECTF curator and, if approved, tagged for submission to NCBI.