

COLLECTF::submission guide (version 2.1)

A play on words using the French *collectif* [collective] and the acronym for transcription factor [TF], COLLECTF is a database of prokaryotic transcription factor binding sites (TFBS). Its main aim 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. The data submitted to COLLECTF gets pushed to the NCBI RefSeq database, where it is embedded as *db_xref* links in complete genome sequences, maximizing the availability of the TF-binding site data and the impact of the research reported by authors.

This document is a companion guide for the submission process. The database is accessible at <http://collectf.umbc.edu>. To read more about COLLECTF, please see the [NAR paper](#) (PMID: [24234444](#)).

COLLECTF::data

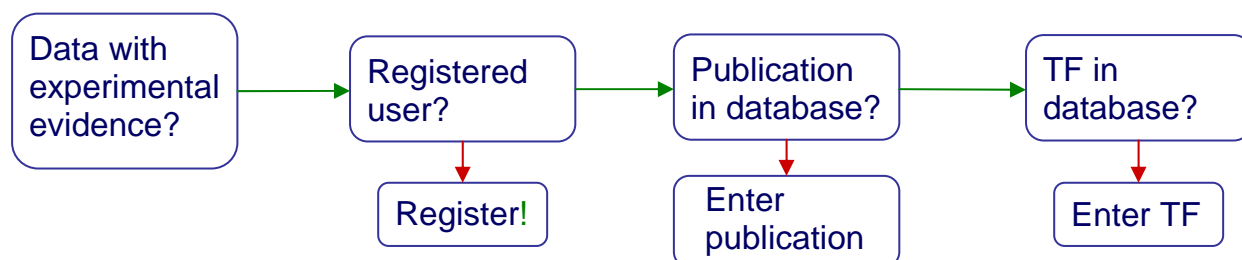
This database only compiles transcription factor binding sites backed by experimental evidence published in peer reviewed articles. COLLECTF distinguishes between two types of experimental support: evidence of binding (e.g. EMSA) and evidence of TF-mediated regulation (e.g. β -gal assay). Identification 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 (not in silico) evidence.*

COLLECTF::before you start

In order to perform a successful submission, several things need to be in place.

User profiles

Before you can submit data to COLLECTF you must first register as a user. To initiate the registration process you must click on the Register link at the upper right of the COLLECTF main page. A valid email address is required for user verification.



Publication submission

Before submitting a curation, the publication that it reports on must be logged in to the COLLECTF database. Please log in and select **New publication** from the **Data submission** menu. You must provide a PMID identifier for your publication and enter name of the transcription factor and species for which the sites are reported. You can indicate, using the appropriate checkboxes, whether your manuscript contains specific promoter information (e.g. Pribnow boxes, transcriptional start site position...) and whether it reports expression data (evidence of TF-mediated regulation).

Publication submission

Use this form to submit a candidate publication for curation in CollectF. Read briefly the publication abstract to validate that it contains relevant data before submitting.

PMID
Paste the PubMed ID obtained from the NCBI website.

Reported TF(s)
Type the name of the transcription factor(s) reported in the manuscript.

Reported species
Type the name of the species reported in the manuscript.

The manuscript contains promoter information
The paper provides experimental data on the structure and sequence of TF-regulated promoter.

The manuscript contains expression information
The paper provides experimental support for TF-mediated regulation of genes.

Submission notes

TF and family information

To submit a curation, you will also need that the TF (and its family) have been added to the database. Please browse the database [by TF family](#) and check whether your specific transcription factor is in the database. If it is not, use the **Add TF** and/or the **Add family** options in **Data submission** to include your TF. You can embed out-links to PubMed and PFAM in the description of TF and family by using the following double colon notation: [PMID::*pmid_accession*] and [PFAM::*pfam_accession*].

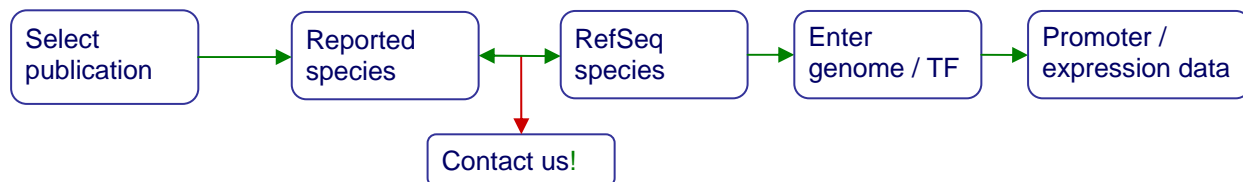
Name

Family

Description

COLLECTTF::curation start

The initial steps of the submission process require that you select a publication and identify a mapping between the species in which you work and available reference genomes in RefSeq.



Step 0: Publication selection

The submission process starts with the submitter selecting a publication for curation. You can upload several publications for curation and perform several curations per publication.

The screenshot shows the COLLECTTF website interface. At the top, there is a navigation bar with links for 'CollecTF', 'About', 'Data submission', 'Admin', and 'More'. On the right, it indicates the user is 'logged in as ivanerill' with a 'Logout' link. Below the navigation bar, a progress indicator shows 'Step 1 of 9'. The main heading is 'Publication selection' with the instruction 'Please choose a publication to curate.' Below this, there is a list of publications under the heading 'Publications'. Two publications are listed, each with a radio button for selection. The first publication is 'The H-NS protein represses transcription of the eItAB operon, which encodes heat-labile enterotoxin in enterotoxigenic Escherichia coli, by binding to regions downstream of the promoter. [15817787]' by Yang J, Tauschek M, Strugnell R, Robins-Browne RM, published in *Microbiology (Reading, England)* 2005 Apr; 151(Pt 4):1199-208. The second publication is 'Reconstruction of the core and extended regulons of global transcription factors. [20661434]' by Dufour YS, Kiley PJ, Donohue TJ, published in *PLoS genetics* 2010 Jul 22; 6(7):e1001027.

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. This is done by providing [RefSeq](#) accession number for the reported chromosomes (e.g. NC_005363.1; including the version number) and TF proteins (e.g. NP_970244; without version number). Notice that RefSeq accession numbers are designated by an underscore; the version number is the one following the period (e.g. NC_005363.1). Only NCBI RefSeq accession numbers are accepted.

Identifying the RefSeq genome matching your experimental species is often a simple step, but it may become complicated if the sequence for the exact strain used in your work is not available as an NCBI RefSeq record. Most often, parental or closely related strains will be available among NCBI RefSeq [genomes](#). As a researcher working hands on with a particular strain, you are best qualified to identify a parental or related strain in NCBI RefSeq.

Nevertheless, if you are uncertain or there is no clear way to identify a surrogate genome in NCBI RefSeq, please [contact](#) the **COLLECTF** team.

Step 2 of 9

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.

Genome NCBI accession number NC_013410.1
Paste the NCBI GenBank genome accession number for the species closest to the reported species/strain. [\[Toggle extra genome accession fields\]](#)

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

TF accession number YP_003250887
Paste the NCBI TF protein accession number for the species closest to the reported species/strain. [\[Toggle extra TF accession fields\]](#)

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

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

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

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. Otherwise, click **This is the same strain...**. This allows us to keep track of the correspondence between reported and mapped strains. You can add more than one chromosome/TF by clicking on **Toggle extra genome accession fields** / **Toggle extra TF accession fields**.

Additional fields

The submission process will ask you to verify again if the manuscript reports promoter information or expression data. Please make sure that **The manuscript contains expression data** is checked if you plan to report differential gene expression associated with TF activity.

- The manuscript contains promoter information
The paper provides experimental data on the structure and sequence of a TF-regulated promoter
- The manuscript contains expression data
The paper provides experimental support for TF-mediated regulation of genes

COLLECTF::curation main

The focus of COLLECTF is on gathering information on experimentally-validated TF-binding sites and mapping it to reference genomes. The main steps of the curation process specifically target these two fundamental points.

Step 2: Experimental methods

Step 2 requires that you report all the techniques used in the paper to verify the TFBS that are being reported in this submission. Most work reporting TF-binding sites involves a heterogeneous mix of techniques (e.g. a site is first shown to bind through footprinting and EMSA, then other sites are validated with EMSA alone). You will be able to specify which technique applies to each site at a later step in the curation process. In this step we also ask that you provide a brief written summary 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. Please provide also external database accession numbers for expression data if applicable (e.g. GEO accession numbers) and, if available, details on whether the TF forms complex with other molecules in order to bind.

Step 2 of 6

Experimental methods used in this paper

Select experimental techniques used to verify binding/expression of the sites reported in the curation. Provide a summary of the basic experimental procedure used to demonstrate binding/expression

Techniques

- 2D PAGE
- Ad-hoc qualitative phenotype observation
- Ad-hoc quantitative phenotype observation
- Alkaline phosphatase reporter assay
- Beta-gal reporter assay
- ChIP-chip
- ChIP-exo
- ChIP-PCR
- ChIP-Seq
- Comparative genomics search

- Western blot (quantitative) expression analysis
- X-ray crystallography
- xyIE reporter assay

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

Experimental process

Experimenters first identified 2 putative binding sites in the flhDC promoter region. Next they ran an EMSA of the promoter region, and found that OmpR bound it. Finally, they used a OmpR::lacZ α operon fusion to perform a β -gal assay which showed a positive regulatory role for OmpR

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

Additional information

- The manuscript reports high-throughput data from an external database. (You can report up to 5 external resources.)
- The manuscript reports that TF forms complex with other proteins for binding with reported sites

External DB type [1]

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

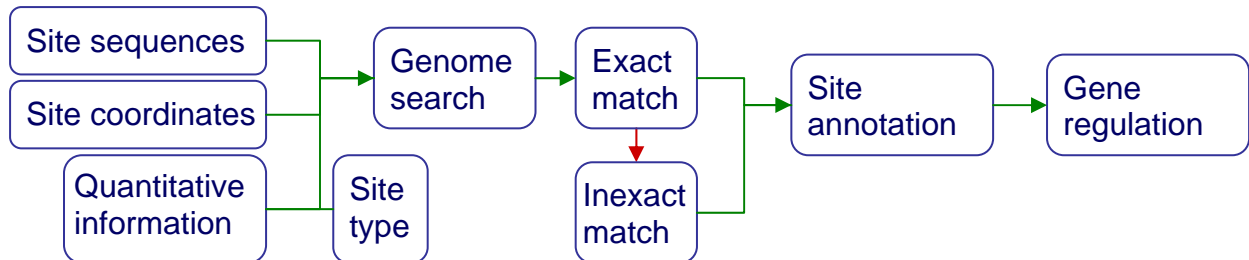
External DB accession number [1]

Type the accession number for external database referenced in paper.

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*". You can browse previous [curations](#) in the database for examples.

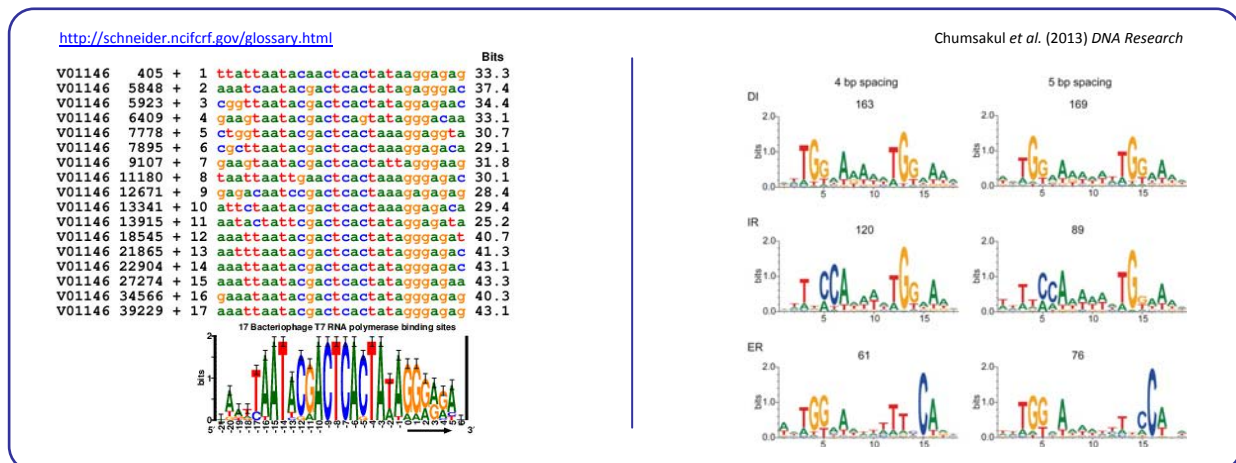
Step 3: Entering reported sites

In this step, you will enter the primary information for **COLLECTF**: binding sites reported in this work *using the techniques specified in Step 2*. Again, you will be able to define what techniques were used specifically for each binding site at a later step.



Site types

TF-binding sites can be defined at 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 (e.g. a ChIP-Seq peak). Many TFs target known specific sequence patterns in the DNA. Some of these patterns are complex and require gapped alignment (e.g. because of variable spacing) or more complex procedures in order to be defined. Other patterns are simpler and can be represented by a gapless alignment of sites (known as a motif), providing a much more concise definition of TF-binding site. In **COLLECTF** we refer to these site types as motif-associated [for gapless alignment], variable motif-associated [for complex patterns] and non-motif associated [for unknown or absent patterns; just evidence of binding]. 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 select either the *Motif associated* or *Variable motif associated* options for your sites. Otherwise, please report them as *Non-motif associated*.



(Left) A collection of “standard” *motif-associated* sites, generating a gapless alignment and typically represented as a sequence logo. **(Right)** Sequence logos for *variable motif-associated* site arrangements of the *Bacillus subtilis* global regulator AbrB, capable of targeting dyads with variable spacer (4-5 bp) and direct or inverted repeats.

Sequence, coordinates and quantitative data

Sites can be entered as sequences (e.g. ATCAGACT) or using genome if they have been mapped to the RefSeq reference strain in the reported work). Sites should be entered one per line (FASTA format is also accepted for sequence entry). In coordinate entry, coordinates are separated by tabs and the first coordinate denotes site start position (e.g. 12280 12260 would denote a 20 bp site in the *reverse* strand starting at position 12280).

If you report quantitative data for sites (e.g. peak intensities, estimated K_d), please append it 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.

Step 4 of 9

Reported sites

Site type	motif associated
Sites	TATGTTAACA 21.5 ATGTTATCGT 12.3 ACTGTTAAGT 11.3 TATGTTCCCTA 12.2 ATTAGTACGT 17.3 TATGTTAGCT 15.1 ATCGTTAACAA 14.4 TCGGTTAGGGT 14.6 ATTATTACGGT 12.2
Quantitative data format	Normalized K_d estimated from quantitative EMSA (Range: 12.2 to 17.3)

Enter the list of sites in FASTA format or type the list of either site sequences or coordinates (one site per line). The sites can be entered in two major formats: sequenced-based (e.g. CTGTTGCAGT) or coordinate-based (e.g. 12312 12323). Optionally, quantitative data (q-val) can also be added to either format. All fields (i.e. site & q-val or coordinates & q-val) must be either space or tab separated.

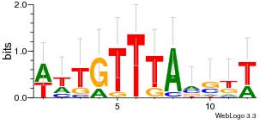
If the manuscript reports quantitative values associated with sites, please enter the quantitative data format here. If not, you can leave this field empty.

Step 4: Verify sites (exact)

Transcription factor binding sites are often submitted as sequences, of which there may be multiple instances in a genome. After submission, sites submitted as sequences must be manually verified by the submitter to validate that the sites entered correspond to a specific genomic location. The COLLECTF submission system will search the genome sequence specified in Step 1 looking for the sequence of each of the sites entered. Exact matches to submitted sites are reported back specifying their location in the genome and nearby genes. Gene annotation details can be accessed by hovering over any gene locus. This information can be used to verify that the sites identified in the NCBI RefSeq genome sequence correspond to the experimentally reported sites.


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.



ATGGTTATCGT

ATGGTTATCGT
+[946336,946347] NC_003210.1



locus tag	gene name	function
Imo0910	Imo0910	Imo0910
Imo0911	Imo0911	Imo0911
Imo0912	Imo0912	similar to transporters (formate)

No valid match.

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 **COLLECTIF** 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.

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.

TATGTTAAACA

TATGTTAAACA
|||||
TATGTTAAACA +[7678,7689] (NC_011186.1)



locus tag	gene name	function
VFMJ11_A0006	VFMJ11_A0006	hypothetical protein
VFMJ11_A0005	VFMJ11_A0005	hypothetical protein
VFMJ11_A0004	VFMJ11_A0004	methyl-accepting chemotaxis protein

TATGTTAAACA
|||||
TATGTTAAACA +[8769,8780] (NC_011186.1)



locus tag	gene name	function
VFMJ11_A0008	VFMJ11_A0008	hypothetical protein
VFMJ11_A0007	sodC_2	copper/zinc superoxide dismutase
VFMJ11_A0009	VFMJ11_A0009	OmpA/MotB domain protein
VFMJ11_A0010	VFMJ11_A0010	Ig domain protein, group 2 domain protein

Step 7: Site annotation

Site annotation step is an essential step for the proper curation of TF-binding site information in COLLECTF. During site annotation, specific experimental techniques are matched to individual sites already identified in reference genome. The quaternary structure of the TF when interacting with sites (e.g. dimer), as well as the regulatory mode of TF-binding at each site (e.g. repressor), if known, can also be entered independently for each site. In addition, if quantitative data for sites has been manually entered or mapped from high-throughput data it can also be validated here. The user can select multiple sites using the mouse in combination with the Shift key or through the `Select/Unselect all` link to easily assign attributes to several sites at once, using the `Apply to selected` option on each column.


Assigning experimental techniques, TF structure or role independently to each site may require some time, but capturing accurate information on the experimental support and nature of TF-binding sites is the main goal of COLLECTF. We therefore kindly request that experimental techniques be completed accurately and that attributes such as quaternary structure be set to default values (`Not specified`) if they cannot be submitted with accuracy. Site annotation can be greatly facilitated by sorting the data before submission, so that sites using similar techniques (or repressed sites, etc.) appear in consecutive order in the `Site Annotation`.

Step 7 of 9

Site Annotation

Fill in the information regarding each site.

Site	TF-type	TF-function	Experimental techniques				Quantitative value
Select/Unselect all	dimer <small>Apply to selected</small>	repressor <small>Apply to selected</small>	Beta-gal reporter assay <small>Apply to selected / Clear all</small>	EMSA <small>Apply to selected / Clear all</small>	qRT-PCR [RNA] <small>Apply to selected / Clear all</small>	Consensus search <small>Apply to selected / Clear all</small>	
<input type="checkbox"/> TATGTTAACA TATGTTGAAAA + [34311, 34322] (NC_000913.2)	dimer	repressor	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	21.5
<input type="checkbox"/> ATGTTTATCGT +[3640020, 3640031] NC_000913.2	dimer	activator	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	12.3
<input type="checkbox"/> ACTGTTTAAAGTT AGTTTGAAGTT +[341, 352] (NC_000913.2)	dimer	repressor	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	11.3
<input type="checkbox"/> TATGTTCCTTA +[535034, 535045] NC_000913.2					<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	12.2



locus tag	gene name	function
b0508	hyl	hydroxypyruvate isomerase
b0509	glxR	tartronate semialdehyde reductase, NADH-dependent

Step 8: 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 8 of 9

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.

The screenshot displays three DNA sequences with their corresponding gene models and checkboxes for selection:

- TATGTTTAAACA**: b0034 (caiF): DNA-binding transcriptional activator. The gene model shows a red vertical bar at the start of the 'caiF' gene.
- ATGGTTTATCGT**: b3496 (dtpB): dipeptide and tripeptide permease B. The gene model shows a red vertical bar at the end of the 'dtpB' gene.
- ACTGTTTAAGTT**: b0002 (thrA): fused aspartokinase I and homoserine dehydrogenase I; b0003 (thrB): homoserine Kinase; b0004 (thrC): threonine synthase. The gene model shows three genes: 'thrA', 'thrB', and 'thrC'.

COLLECTF::curation end

The final step in the COLLECTF submission pipeline deals with validating the curation, adding notes and determining whether curation is complete for the original manuscript.

Step 9: 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 was identified 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 should be considered for submission to NCBI. Curations will only be considered for submission to NCBI if the sequence for the 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*.

Multiple curations

The system also requires that you specify whether the Curation for this paper is complete. Do not check this box if, for instance, you want to report additional sites, regulatory modes and/or sources of experimental support in a subsequent curation, or if you are reporting data for more than one TF or species. The COLLECTF submission system allows you to submit data from a literature source in as many independent submissions as you require in order to facilitate the Site Annotation step in each submission. The submission system will pre-populate fields in subsequent submissions, so that only reported sites and their annotation must be entered anew in each submission (all other fields can, but do not *have to*, be edited). The same sites can be submitted multiple times (e.g. with different experimental evidence). The COLLECTF system will automatically integrate all the data reported for one site.

Revision required

When no genome remotely resembling that of the reported species is available in RefSeq, if sequencing of the genome is still in progress or if the TF of interest is not available in RefSeq, the submission should be tagged as requiring revision. The data for submissions requiring revision is stored in the database, and the COLLECTF team periodically assesses whether the conditions for revision are met in order to finalize the submission and link it to RefSeq records.

Final submission

After you check I want to submit this curation and click Next, a summary of your submission will appear for your review. If you spot any errors in the submission, please let us [know](#) immediately.

Step 9 of 9

Curation information

This step finalizes the curation. Fill all required fields.

Revision required None

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

Curation for this paper is complete.
Check this box if there are no more curations pending for this paper (additional sites, sites supported by different techniques, sites for other TFs, etc).

Notes
sites for AbrC will be reported in a separate curation

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 tagged for submission to NCBI. On behalf of the COLLECTF team, THANK YOU for your contribution!