

Worked-out examples

This guide aims to provide a series of example sentences which ECO curators have previously come across, coupled with in-depth descriptions of what annotations were (or were not) made and why. We include some simple examples of correct and incorrect annotations designed to point out common signs of good ECO annotations. And we show and explain some more difficult annotations or fringe cases which we have come across in our research to help curators navigate the often challenging aspects of parsing research papers at a sentence level.

These different examples are drawn from various manuscripts and explain the reasoning behind why different annotations were (or not) made. Each example will be marked with its Pubmed Identification Number (PMID), an 8-digit code which curators use to concisely identify papers.

Part 1: Examples of Sentences to Annotate

Detailed here are a series of simple, cut-and-dry style annotations containing a clear listing of both **evidence** (usually a wet-lab or dry-lab method or technique) and **assertion**. An annotation will always consist of these two fundamental parts, but some sentences have more clearly-defined components than others. An important point is that "evidence" is a technique or method that is being used to make this assertion. A statement that only contains a technique or method is not truly evidence and so will not be annotated.

Examples with single evidence occurrence and single assertion occurrence

"The RT-PCR assay indicated that the *sycO*, *ypkA* and *yopJ* genes (designated as pCD12, pCD13 and pCD14 in *Y. pestis* 91001 [19], respectively) were transcribed as a single primary RNA (Fig. 1), and thereby these three genes constituted a single operon in *Y. pestis* Microtus strain 201."

Paper: "Direct and negative regulation of the *sycO*-*ypkA*-*yopJ* operon by cyclic AMP receptor protein (CRP) in *Yersinia pestis*."

PMID: 19703315

This is a good example of an annotation, with a clear statement of method and a clear result detailing what the researchers used this method to discover. In this sentence, these researchers are using Reverse-Transcription Polymerase Chain Reaction (RT-PCR) to determine that three genes belong to an operon in this species.

We can annotate the sentence like this:

- [ECO: 0000109](#)
- **Name:** Reverse-Transcription Polymerase Chain Reaction Evidence
- **Term Confidence:** High
The sentence directly states that RT-PCR was used to inform their assertion.
- **Assertion Strength:** High
The sentence affirmatively states its assertion as factual.
- **Category:** Sequence Feature (SO)
The assertion is about an operon, a sequence feature.
- **Sentence Pair:** No
- **Negative Assertion:** No

"Phylogenetic analysis of the L. helveticus CM4 BCARR protein revealed the presence of homologs in lactobacillaceae, enterococcaceae, leuconostocaceae, carnobacteriaceae, listeriaceae, exiguibacteria, and bacillaceae (Figure 8)."
PMID: 24146802

This sentence has one statement of evidence (phylogenetic analysis) and one assertion about the homologs of CM4 BCARR in the various bacteria families.

- ECO:0000080
- **Name:** Phylogenetic Evidence
- **Term Confidence:** High
From "Phylogenetic evidence".
- **Assertion Strength:** High
From "revealed the presence of".
- **Category:** Taxonomy/Phylogeny
From "homologs in".
- **Sentence Pair:** No
- **Negative Assertion:** No

"Northern blot analysis indicated that there was one major band of 1.5-kb (which was used for the stability determination) and two other minor bands of approximately 3.0- and 6.0-kb, which constituted less than 5% of the total signal (Fig. 1C), suggesting that other genes may be co-transcribed along with SMU.1882."

Paper: "Activation of the SMU.1882 transcription by CovR in Streptococcus mutans."
PMID: 21124877

This sentence contains a good example of a clear but low-confidence assertion. While our method here is of very high confidence (northern blot analysis), our assertion is clearly present but heavily modified to assuage confidence. The usage of "suggests" and "may" make this sentence's conclusion much weaker. This isn't necessarily a bad thing— some scientific tests simply are less trustworthy than others when it comes to certain kinds of conclusions, and a good researcher does not want to make an unsubstantiated claim.

With that in mind, here's how we annotate this sentence:

- [ECO:0000106](#)
- **Name:** Northern Blot Evidence
- **Term Confidence:** High
The researchers clearly state their technique and that it produced their results.
- **Assertion Strength:** Low
The sentence contains a clear assertion, but it is heavily modified to assuage confidence in the technique's findings, with the words "suggesting" and "may be".
- **Category:** Sequence Feature (SO)
IMPORTANT POINT: *This sentence talks about transcription, other genes being co-transcribed with the product. This cannot be annotated with the category Molecular Function, Biological Process, or Cellular Component as there is **no** gene-product object of assertion, but it does fall under the category of sequence feature.*
- **Sentence Pair:** No
- **Negative Assertion:** No

"Previous experimental evidence suggests that all four of these hrp promoters are genuine."

Paper: ."
PMID: 25170934

This sentence is a pretty tricky one to annotate. While the actual experimental method that "previous experimental evidence" implies is unknown to us, this sentence still contains both halves of an evidence statement: a method and an assertion. The phrase "experimental evidence" is an ECO term and so is relevant for annotation. Also note that when a paper references the work that another research team has done, we can still annotate it, but we must be cautious when doing so to not assume any outside information. Sometimes whether to annotate these kinds of high-level experimental statements is a judgment call based on the rest of the sentence.

Here's how we'd annotate this:

- [ECO:0000006](#)
- **Name:** Experimental Evidence
- **Term Confidence:** High
While we know this sentence does very clearly contain a type of evidence, we do not know what evidence type there is. Therefore, we must use the simplest term possible which still conveys meaning.
- **Assertion Strength:** *Medium*
This is a little bit of semantics. The term "suggests" generally carries a less definite meaning than a more conclusive word like "established", "confirmed", or "Indicated", so while the assertion is definitively present, medium confidence is safe.
- **Category:** Sequence Feature
This sentence is talking about the identity of a promoter, so it falls under the category of sequence feature.
- **Sentence Pair:** No
- **Negative Assertion:** No

Binding is a tricky annotation

Although we want our assertions to be very clear -- "and therefore, we show that X binds Y", for binding statements this is not always the case. And yet, a sentence without such a clear assertion could clearly mean that binding occurs. Here is one such example -- a "stable

complex" between a protein and a DNA region means that binding occurred. "Binding" is one of the few types of assertions where statements without a definitive "and therefore", "thus", "and this evidences suggests that ..." or similar phrase can be considered a clear enough assertion to annotate.

"The electromobility shift assay shows that MelR forms more stable complexes at the TB22 promoter than at the TB28 promoter (Figure 7)."

Paper "Autoregulation of the Escherichia coli melR promoter: repression involves four molecules of MelR"

PMID: 18346968

In this sentence, researchers are using an electrophoretic mobility shift assay (EMSA) to test the binding affinity of the protein MelR to two possible DNA probes.

Here's how we annotate this sentence:

- [ECO: 0000096](#)
- **Name:** Electrophoretic Mobility Shift Assay Evidence
- **Term Confidence:** High
The sentence clearly states that EMSA was used to find the information presented in the sentence.
- **Assertion Strength:** Medium
The sentence states a clear assertion, but due to the comparison used in the sentence we don't entirely have the context to understand the conclusion (remember- we cannot use context from other parts of a paper to analyze a single sentence). So, we can mark this as medium.
- **Category:** Molecular Function (GO)
EMSA is used to assay a molecule's ability to bind (usually to DNA), and protein binding falls under the Gene Ontology's Molecular Function category. From "stable complexes".

Example of backwards assertion

"To detect the binding of FNR to the predicted target promoters in vitro, the radiolabelled DNA fragments were used in EMSA assays with purified FNR protein (Figure 2B). In all cases the addition of purified FNR retarded the migration of the purified DNA fragments."

PMID:17164287

The assertion statement is setup in the **first** sentence (not the second as is normally the case) as to what is being assessed ('To detect the binding of FNR'), and then the readout **confirms** the statement -- the second statement clearly states the positive result of the experiment. This situation makes the pair an 'assertion' in the sense that no inference is needed to make a molecular function annotation out of this sentence pair.

- ECO:0000096
- **Name:** Electrophoretic Mobility Shift Assay Evidence
- **Term Confidence:** High
From "EMSA assays".
- **Assertion Strength:** High

The second sentence makes a clear statement of what was observed "in all cases, addition of purified FNR retarded the migration". There is no "seemed to retard" or similar weakening of the assertion.

- **Category:** Molecular Function
From "the binding of FNR to the predicted target promoters". Binding is normally a Molecular Function.
- **Sentence Pair:** No
- **Negative Assertion:** No

Example of wrap up sentence that can be annotated

"These data, coupled with the microarray results, suggest that PreA is necessary for the activation of the ygiW-STM3175, preA-preB, and mdaB-yglN operons."
PMID:19236707

This is a wrap up sentence. Normally wrap up sentences rely on evidence from more than just the previous sentence and are not annotated. This sentence partially relies on previous sentences ("these data"), but notice that part of the evidence, the microarray results, is repeated in the sentence. So this evidence is right here in the sentence, and the annotation can be made for that.

- ECO:0000058
- **Name:** Expression microarray evidence
- **Term Confidence:** Medium or low
From "microarray results". There are several microarray ECO terms so we can't be highly confident just based on "microarray results".
- **Assertion Strength:** Medium
The assertion is from "suggest".
- **Category:** Biological Process
From "PreA is necessary for the activation of...". This is a synonym for positive regulation, a Biological Process.
- **Sentence Pair:** No
- **Negative Assertion:** No

Example with more than one evidence statement but only one assertion

"The resemblance of the candidate promoters to the canonical hrp promoter consensus sequence, together with the evidence that HrpL binds at their genomic locations suggests that they are genuine HrpL-responsive promoters."

Paper
PMID: 25170934

Here, we have a two-part sentence, with two pieces of evidence being used to make an assertion. One interesting aspect of this sentence is that both techniques share the same assertion statement, namely "suggests that they are genuine HrpL-responsive promoters."

First, we have this part of the sentence:

"The resemblance of the candidate promoters to the canonical hrp promoter consensus sequence..."

We annotate this sentence like so:

- [ECO:0005532](#)
- **Name:** Consensus Search Evidence
- **Term Confidence:** High
"Resemblance of the candidate promoters" and "hrp promoter consensus sequence" tip us off to this sentence being consensus search evidence.
- **Assertion Strength:** Medium
"Suggests that they are genuine" implies that the sentence's techniques generated some sort of output or result.
- **Category:** Sequence Feature
This sentence is making an assertion about the existence of promoters ("genuine HrpL-responsive promoters"), so it falls under the category of sequence feature.
- **Sentence Pair:** No
- **Negative Assertion:** No

Next up we have:

"...together with the evidence that HrpL binds at their genomic locations..."

This section can be annotated this way:

- [ECO:0000024](#)
- **Name:** Protein Binding Evidence
- **Term Confidence:** Medium
"binds at their genomic locations" implies that protein binding evidence is being done in this sentence, but we cannot be certain because without outside context we do not know that HrpL is a protein. Similarly here, we must remember that machines do not possess reading comprehension skills like human curators do.
- **Assertion Strength:** Medium
- **Category:** Sequence Feature
This sentence fragment is providing evidence for the same assertion as above, namely the HrpL-responsive promoters.
- **Sentence Pair:** No
- **Negative Assertion:** No

Example with multiple evidence and assertion statements

The footprinting and EMSA data show that different nucleoprotein complexes are formed at PhlyE depending on the relative concentrations of H-NS and SlyA through competition for overlapping binding sites upstream and downstream of the hlyE transcript start site. This suggested a regulatory mechanism whereby increased

intracellular concentrations of SlyA remodel H-NS binding at PhlyE to relieve H-NS-mediated repression by allowing RNAP to access the promoter.
PMID:17892462

The above sentence pair contains two assertions in the first sentence and two separate ones in the second sentence. Note that the evidence for the second sentence is found in the first one, and that the **same** evidence in the first sentence is used for multiple assertions in both sentences. Finally, there is multiple evidence too.

There are **four** annotations for the **first** sentence alone because both evidence types are used to make two assertions in this sentence. Note that in brat **all 4** annotations should be entered because each one has a unique combination of ECO identifier and category.

- ECO:0000096
- **Name:** Electrophoretic Mobility Shift Assay Evidence
- **Term Confidence:** High
From "EMSA".
- **Assertion Strength:** High
From "show"
- **Category:** Molecular Function
From "nucleoprotein complexes are formed". A nucleoprotein complex is protein bound to DNA, so it is binding which is normally a Molecular Function.
- **Sentence Pair:** No -- Note the NO. Here the annotation is for the first sentence only.
- **Negative Assertion:** No
- ECO:0000136
- **Name:** Nucleic Acid Binding Evidence
- **Term Confidence:** Low
From "footprinting".
- **Assertion Strength:** High
From "show"
- **Category:** Molecular Function
From "nucleoprotein complexes are formed". A nucleoprotein complex is protein bound to DNA, so it is binding which is normally a Molecular Function.
- **Sentence Pair:** No -- Note the NO. Here the annotation is for the first sentence only.
- **Negative Assertion:** No
- ECO:0000096
- **Name:** Electrophoretic Mobility Shift Assay Evidence
- **Term Confidence:** High
From "EMSA".
- **Assertion Strength:** High
From "show"
- **Category:** Sequence Feature
From "overlapping binding sites".
- **Sentence Pair:** No -- Note the NO. Here the annotation is for the first sentence only.
- **Negative Assertion:** No
- ECO:0000136
- **Name:** Nucleic Acid Binding Evidence
- **Term Confidence:** Low
From "footprinting".

- **Assertion Strength:** High
From "show"
- **Category:** Sequence Feature
From "overlapping binding sites".
- **Sentence Pair:** No -- Note the NO. Here the annotation is for the first sentence only.
- **Negative Assertion:** No

There are also 4 annotations for the sentence **pair** with the evidence in sentence one (same evidence as above) for the two assertions in sentence two. The tie between the two sentences is "This" (and the fact the two sentences are consecutive). Here these annotations are sentence pairs.

BUT note that in brat **only 2** annotations should be entered because both assertions are for the same category, Biological Process. So there are only two unique combinations of ECO identifier and category here. However, all four are listed below for clarity of seeing which words correspond to which annotation.

- ECO:0000096
- **Name:** Electrophoretic Mobility Shift Assay Evidence
- **Term Confidence:** High
From "EMSA".
- **Assertion Strength:** Medium
From "suggested"
- **Category:** Biological Process
From "allowing RNAP to access the promoter". This means the regulation of the RNA binding transcription factor. Regulation is a biological process.
- **Sentence Pair:** Yes -- sentence 1 and 2 must be paired to form this assertion.
- **Negative Assertion:** No
- ECO:0000136
- **Name:** Nucleic Acid Binding Evidence
- **Term Confidence:** Low
From "footprinting".
- **Assertion Strength:** High
From "show"
- **Category:** Biological Process
From "allowing RNAP to access the promoter". This deals with regulation of RNA binding transcription factor activity. Regulation is a biological process.
- **Sentence Pair:** Yes -- sentence 1 and 2 must be paired to form this assertion.
- **Negative Assertion:** No
- ECO:0000096
- **Name:** Electrophoretic Mobility Shift Assay Evidence
- **Term Confidence:** High
From "EMSA".
- **Assertion Strength:** High
From "show"
- **Category:** Biological Process
From "a regulatory mechanism... to relieve H-NS-mediated repression"
- **Sentence Pair:** Yes -- sentence 1 and 2 must be paired to form this assertion.

- **Negative Assertion:** No
- ECO:0000136
- **Name:** Nucleic Acid Binding Evidence
- **Term Confidence:** Low
From "footprinting".
- **Assertion Strength:** High
From "show"
- **Category:** Biological Process
From "a regulatory mechanism... to relieve H-NS-mediated repression"
- **Sentence Pair:** Yes -- sentence 1 and 2 must be paired to form this assertion.
- **Negative Assertion:** No

Part 2: Examples that are NOT annotated

"The resulting RNA-Seq data contain both whole transcriptome and TSS information (Table 2)."

This is not an evidence sentence. While it contains direct references to RNA sequencing, which is an experimental technique, **there is no assertion being made** in this sentence. All it tells us is that the data contains information, not what that data implies about the experiment being performed.

"We extracted 50 nucleotides directly upstream from each captured 59-end, resulting in 1451 sequences derived from the (delta)hrpL-FLAG sample and 1472 sequences from the hrpL sample (overlapping sequences within a sample were merged) and used the sequences as input to MEME [51]."

Similar to the previous example, this is not an evidence sentence, though it is a bit less straightforward, as it contains words like "used" and "extracted" which make it sound like the sentence is generating some sort of result. However, close inspection reveals that the sentence **does not actually present any evidence**, it only dictates an experimental technique being conducted by the researchers.

"Validation tests suggest that this promoter supports HrpL-dependent transcription (below)."

At first glance, this sentence appears to be a convincing evidence statement, containing a technique, albeit a vague one, in "validation tests", as well as words like "suggest" and "supports" which would make the sentence appear to contain evidence. However, **we cannot annotate this sentence** because understanding the technique used requires us to utilize evidence in a different part of the paper. "Below" here refers us to the next paragraph of the paper, where the validation tests, a reporter gene fusion assay and ChIP-qPCR, are described in detail.

This sentence is an example of what we call a *cataphoric reference*, a sentence which refers the reader to information located later on in the paper. As our eventual goal is to create a machine learning system using our corpus of annotated papers, we must leave these sentences unannotated, as our learning system will not be able to understand the reference to other areas of the paper. In order to avoid *false positives* such as this sentence, it is important to remember that we cannot use any context outside of the sentence we are annotating to make a call on an evidence statement.

This sentence differs also from the example that had "experimental evidence". Without knowing the rest of the paper, "Validation tests" could mean anything -- a wet lab experiment, a dry lab experiment, someone manually checking results (that is not an "experiment"). So without more information we can't select an ECO term for this one.

"Sequence pattern matching has been used extensively to inventory the HrpL regulon in DC3000."

This sentence is another example of a very convincing false positive. While the sentence appears to contain an evidence type ("sequence pattern matching") and an assertion ("to inventory the HrpL regulon"), a closer look shows that **the assertion here does not actually signify that any evidence has been discovered**. The "has been used" part of the sentence is extremely vague and changes the sentence's meaning from "we used sequence pattern matching to inventory the regulon" to "Sequence pattern matching is used to inventory the regulon." This is more a case of the author saying **why** a scientific technique is used, not what evidence they generated using the technique. Therefore, we do not annotate this sentence. If the sentence had included a clause that said something like "and therefore, we determined that HrpL controls the transcription of genes X, Y, and Z", it would have an assertion that could be annotated.

Keep in mind that normally if a sentence says something like "to test", "in order to test", "to show", "to investigate", "in order to determine", etc., that the sentence is stating the purpose but is not making an assertion.

"As expected, the ompF promoter activity (beta-galactosidase activity) decreased significantly in DeltaompR relative to WT grown at high medium osmolarity (0.5 M sorbitol); however, it showed almost no difference between WT and C-ompR, thereby confirming that the ompR mutation was nonpolar."

PMID: 21345178

This is a perfect assertion with evidence BUT the assertion (about the mutation being nonpolar) is **NOT one of our categories**. So no annotation can be made. If you are unsure about whether an assertion is about one of the categories -- please do ask.

"We found that compared to that of wild type, toxR-lacZ expression was reduced in aphB mutants, while expression of aphB from a plasmid in this mutant restored toxR expression (Fig. 4B) and ToxR production (Fig. 4C)."

From "Virulence regulator AphB enhances toxR transcription in *Vibrio cholerae*."
PMID: 20053280

The sentence contains an example of common evidence: mutant phenotype evidence (ECO:0000015). Compared to some other experimental techniques, mutant phenotype evidence tends to be a little difficult to spot for first-time annotators because it often does not contain some of the common terms that other types of annotations do. For a sentence to be considered high confidence mutant phenotype evidence, it must contain an instance of a mutant organism or gene product being used and a comparison of that mutant product to another phenotype (either a wild type or another mutant). The sentence also mentions "-lacZ" which is a clear reference to the beta-galactosidase reporter (ECO:0000096).

However, there is **no** clearly stated assertion. Instead, the use of the mutant and the change in the expression **IMPLIES** that the gene aphB regulates transcription. But the sentence doesn't clearly say so, thus no Biological Process annotation can be made.

Also, because a gene (aphB) is clearly being discussed in the sentence, we cannot annotate with the category Phenotype.

So no annotation will be made for the above sentence. It will be considered a readout.

"Although the scan matched all annotated and new candidate hrp promoters identified in this study, the model did not match any other region in the genome that showed enrichment in the ChIP-Seq experiment (Evalue cut-off = 0.001, 245 promoter candidates in total)."

This sentence mentions the ChIP-Seq experiment but all it is talking about is a scan matching various promoters without explicitly saying which promoters. Thus it is too vague to annotate.

Sentence #1: "The stacking energy profiles of R.etli and E.coli promoter regions were variable, but with a tendency to low negative values (low stability), nevertheless local minimum values were located around the -10 box."

Sentence #2: "In contrast, the stacking energy profiles of R.etli and E.coli coding regions were similar: both showed more negative values that corresponded to great stability (Figure 2a and b)."

Sentence #3: "These results suggest that despite the variability of the nucleotide composition of the R.etli promoters, these regions

possess thermodynamic and structural properties similar to the E.coli promoter regions."

The final sentence in the above set is a wrap up sentence since it is relying on sentence #1 and #2 for evidence (stacking energy) (most specifically it relies on sentence #1 which is too far away). Furthermore, in this case, the assertion is about "thermodynamic and structural properties", which is not one of our categories. Therefore, there would be no annotation here.

Part 3. Sentences that have some annotatable parts and some not

"We first confirmed that cells bearing the tagged protein retained the ability to stimulate the hypersensitive response in a plant assay (Figure 1B) and established that HrpL-FLAG retained its ability to support transcription from a known HrpL-responsive promoter (Figure 1C)."

This is a two-part sentence, referring to two different techniques which are implicated in the generation of evidence.

The first part of the sentence reads:

"We first confirmed that cells bearing the tagged protein retained the ability to stimulate the hypersensitive response in a plant assay (Figure 1B)..."

We can tag this sentence in the following way:

- [ECO: 0000059](#)
- **Name:** *Experimental Phenotypic Evidence*
- **Term Confidence:** *High*
The use of "cells bearing the tagged protein" (ie, the expression of a phenotype) justifies the use of ECO:0000059.
- **Assertion Strength:** *High*
"Confirmed" generally implies a strong assertion, as it shows that the technique used in the sentence produced a result.
- **Category:** *Phenotype*
"Stimulate the hypersensitive response" implies that a phenotypic change is being measured. Therefore, this falls under the classification of phenotype.

Meanwhile, the second part of the sentence states:

"...and established that HrpL-FLAG retained its ability to support transcription from a known HrpL-responsive promoter (Figure 1C)."

This part of the sentence does not mention evidence relevant for the assertion about "supporting transcription". Thus, no annotation for this portion of the sentence will be created.