



# Reannotation of Prokaryotic Genomes

The introduction of non-redundant annotation reagents, WP\_ accessioned RefSeq proteins  
<http://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/reannotation/>

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

## What we are doing

### **Prokaryotic RefSeq genome re-annotation using an improved prokaryotic annotation pipeline**

All new WGS and complete genomes and all pre-existing RefSeq WGS and complete genomes are being re-annotated by an updated prokaryotic annotation pipeline. Concurrent with this update, we are managing gene and protein data produced from annotation of prokaryotic RefSeq genomes with a new non-redundant protein data model including the RefSeq protein accessions with prefix "WP\_" as documented.

Announcement on WP <http://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins>

Note that a small number of high quality reference genomes that have been extensively curated, community-vetted and/or have a large amount of experimental confirmation will retain their existing annotation. For example, when we have information about an important representative genome, such as *Escherichia coli* K12 substr. MG1655, it will continue to be annotated with species and strain-specific RefSeq proteins with "NP\_" or "YP\_" accession prefixes, e.g.

**NP\_414543.1**. However, protein coding regions that have been annotated in other non-reference genome strains, such as *Escherichia coli* HM605, will be cross-referenced with protein records represented by the "WP\_" non-redundant protein accession prefix, e.g. **WP\_001264707.1**. Direct URLs for these reference records are listed below.

A representative genome *E. coli* K12 substr. MG1655 [http://www.ncbi.nlm.nih.gov/nuccore/NC\\_000913.3](http://www.ncbi.nlm.nih.gov/nuccore/NC_000913.3)

A protein annotated on representative genome [http://www.ncbi.nlm.nih.gov/protein/NP\\_414543.1](http://www.ncbi.nlm.nih.gov/protein/NP_414543.1)

A non-reference genome *E. coli* HM605 [http://www.ncbi.nlm.nih.gov/nuccore/NZ\\_AJWU01000012.1](http://www.ncbi.nlm.nih.gov/nuccore/NZ_AJWU01000012.1)

A protein annotated on non-reference genome [http://www.ncbi.nlm.nih.gov/protein/WP\\_001264707.1](http://www.ncbi.nlm.nih.gov/protein/WP_001264707.1)

## Why are we doing this?

### **Reduced protein redundancy and improved genome annotation**

We are now receiving very large numbers of nearly identical bacterial genome sequences from environmental and clinical samples as well as those isolated for identification of food borne pathogens. Annotating the genes on these genomes is important and useful for identifying the small number of places where mutations may affect function, but most of the genes annotated on these new genomes encode proteins that are identical to proteins already in the RefSeq dataset. This would produce a large number of redundant records without improving the usability and usefulness of this information. Changes in the prokaryotic protein data model are designed to address this issue by creating a new data model for identical protein sequences produced by the prokaryotic genome annotation pipeline.

Managing quantity and quality of annotated gene and protein data in NCBI's Nucleotide, Protein and Gene databases is critical for the application of prokaryotic genomics data in health-related use cases. Therefore, in order to continue to provide a prokaryotic dataset that is of highest utility for disease, pathogen, and other comparative analysis needs, NCBI has decided to re-annotate all RefSeq prokaryotic genomes using an improved genome annotation pipeline which features a consistent method for gene annotation with improved management of protein names.

## What database records are impacted by this?

### **Nucleotide, Protein & Gene database records**

Due to the increase in submissions of genome sequences and subsequent generation of a large volume of identical gene annotations and protein sequences, a change was necessary in how the prokaryotic genome annotation pipeline produces records for the NCBI Protein and NCBI Gene databases, as well as corresponding cross-references noted on NCBI Nucleotide records.

NCBI has introduced a new data model for prokaryotic RefSeq genomes, and a new protein data type in the RefSeq collection that is signified by a "WP\_" accession prefix. If the identical protein sequence (exactly the same protein sequence and length) appears on more than one RefSeq genome, NCBI re-uses the existing "WP\_" accession instead of generating a new one for each new occurrence. For conserved proteins the same "WP\_" accession may appear on thousands of genomes.

In addition, the scope of Gene for prokaryotes has been changed from including data from all complete genomes, to incorporating only those that are: a) a reference genome; or b) a representative genome for which there are at least 10 sequenced genomes for the species or clade. As such, the Gene database will now include only the highest quality and most supported subset of RefSeq prokaryotic genomes.

## Description of key data changes and how to transition to the new data

**New genomic annotation data model:** The new data model is now in use for both WGS and Complete bacterial genomes. Archaeal RefSeq genomes are also using this data model but the transition has not been completed as of May 2015.

**Re-annotated genomes:** Over 32,000 prokaryotic RefSeq genomes have now been annotated using NCBI's improved prokaryotic genome annotation pipeline.

**Reference genomes:** 122 bacterial reference genomes have been selected based on genome and annotation quality, established community reference standards, and experimentally supported community annotation as delineated in the selection criteria:

<http://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/#referencegenome>

**Representative genomes:** 3,152 representative genomes have been selected based on considerations that include genome quality, annotation quality, and comparison to similar genomes after clustering clades. More details are at:

<http://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/#representativegenome>

**Suppressed genomes:** Several hundred RefSeq bacterial genomes that did not pass assembly or annotation quality validation have been suppressed. A comprehensive list of suppressed Nucleotide genome accessions is available. In addition, a report for a set of recently suppressed genomes is available separately. These two files are at:

<ftp://ftp.ncbi.nlm.nih.gov/refseq/release/release-catalog/release70.removed-records.gz>

<ftp://ftp.ncbi.nlm.nih.gov/refseq/release/release-catalog/release70.addedQA-suppressedAssemblies.txt>

**Suppressed proteins:** Approximately 7 million prokaryotic NP\_ and YP\_ accessioned proteins have been suppressed as a result of both suppressing RefSeq genomes with quality issues and due to the large-scale re-annotation project that transitioned bacterial genomes to the new non-redundant protein data model. An informational message now appears on suppressed sequences that are identical to a non-redundant RefSeq protein record. An example with informational links (A) is shown to the right. Follow the "Re-annotation project" link for more information.

### hypothetical protein SYNGTI\_1741 [Synechocystis sp. PCC 6803 substr. GT-I]

⚠ Reference sequence YP\_005383503 has been replaced by WP\_010872936.1

The sequence YP\_005383503 is 100% identical to WP\_010872936.1 over its full length. Be aware that a NCBI nonredundant RefSeq protein (WP\_) can be annotated on large numbers of bacterial genomes that encode that identical protein.

[Old YP\\_005383503.1](#) [New WP\\_010872936.1](#) [Identical proteins](#) [Re-annotation project](#)

[http://www.ncbi.nlm.nih.gov/protein/YP\\_005383503.1](http://www.ncbi.nlm.nih.gov/protein/YP_005383503.1)

**Changes in Gene:** The scope for production of prokaryotic genomes with NCBI GeneIDs has changed, resulting in the suppression of a large number of entries. An informational message now appears on suppressed Gene database records. An example display with relevant

### rpsI 30S ribosomal protein S9 [Pseudomonas fluorescens A506]

Gene ID: 12962335, discontinued on 4-Mar-2015

⚠ All Gene records for this genome have been discontinued due to a change in scope for [prokaryotic genomes in Gene](#). At the time this Gene record was discontinued, the RefSeq genome was re-annotated with the following features:

locus\_tag: PFLA506\_RS04145

protein: [WP\\_002555064.1](#)

location: [NC\\_017911.1 \(955969..956361\)](#)

annotation change: identical

<http://www.ncbi.nlm.nih.gov/gene/12962335>

links (B) to current protein record and genomic subsequence is shown above.

**Locus\_tag changes:** New locus\_tags are assigned as genomes are re-annotated using the format `<original locus_tag prefix>_RS<digits>` (for example: MAP4\_RS09710). The previously annotated locus\_tag will be maintained on the Nucleotide record but will now be reported in the 'old\_locus\_tag' qualifier when the equivalent CDS has been re-annotated on the record.

### Mapping file

Suppressed records with "NP\_" or "YP\_" accessions, NCBI GeneIDs, and locus\_tags are mapped to updated RefSeq data here:

<ftp://ftp.ncbi.nlm.nih.gov/refseq/release/release-catalog/release70.bacterial-reannotation-report.txt.gz>

## Learn More and Keep Up-to-Date

Read more about this <http://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/reannotation/>

RefSeq policy and data content announcements <http://www.ncbi.nlm.nih.gov/mailman/listinfo/refseq-announce>

Other NCBI-wide announcements:

NCBI News <http://www.ncbi.nlm.nih.gov/news/>; NCBI Twitter account <https://twitter.com/ncbi>