New RefSeq protein product and data model

Background

With the advent of high-throughput sequencing, the generation and use of genome sequences is evolving. One such change is the sequencing of very large numbers of nearly identical bacterial genomes to analyze food borne pathogens or infection outbreaks. In these cases, 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 one already in the RefSeq dataset.

In order to manage the flood of identical proteins and decrease existing redundancy, particularly from bacterial genomes but soon from viruses and eukaryotes as well, NCBI is introducing a new protein data type in the RefSeq collection signified by a 'WP' accession prefix. WP accessions provide non-redundant identifiers for protein sequences. This new data type is provided through NCBI's genome annotation pipeline but will be managed independently of the genome sequence data to ensure the dataset remains non-redundant. We are doing this for two major reasons: 1) WP protein records represent a non-redundant protein collection that provides information about the protein sequence and name with linked information to genomic context and taxonomic sample; 2) use of WP accessions allows us to avoid creating millions of redundant protein records in the RefSeq collection.

When the NCBI genome annotation pipeline annotates a bacterial protein that is 100% identical and the same length as an existing WP accessioned protein (or is truncated due to an assembly gap) NCBI will no longer create a new protein record (with exception below). Instead, NCBI will annotate that protein on the genome by referencing the existing WP accession in the annotated CDS feature, indicating that the genome represents yet another exact example of that known protein sequence. Any annotation of protein function on the genome record (such as the product name) will be inherited from the independent WP record. WP records, therefore, always mean one exact sequence that may be observed only once or many times in different strains or species. Also, WP records will always have a version of "1" and will not be updated like taxon-specific RefSeq records.

Where an important representative genome exists, such as *Escherichia coli K12*, that has an existing set of annotated proteins with NP or YP accessions, the representative genome will retain its current protein accessions; however, these protein records will refer to the matching WP protein accession in the sequence block. This approach allows the NP (or YP) protein records to continue serving their function of defining taxonomically-oriented sets of proteins and to track sequence changes over time. Each NP will continue to carry its authoritative annotation, single taxonomic name, revision history, and version number (that will continue to update if the sequence is revised). If a NP protein sequence is revised by curation, and its version number changes, it will now refer to a different WP record that is identical to the updated NP sequence. For some species NCBI will provide more than one representative genome that, for example, reflects existing community standards or has specific experimental support.

Obviously this is too strict a rule to group non-identical but similar proteins which have similar or even identical function. So there will be a (future) process of clustering WP proteins into functional groups, and assigning separate identifiers and names for such groups. WP records within the same cluster may indicate sequence differences resulting from population variation or from incomplete genome sequence data; protein clusters may therefore indicate a larger group of related genomes than any individual WP record. Thus, this data model not only reduces the redundancy in the protein dataset, it also removes redundancy from clustering and other approaches that use the WP protein dataset.

Details:

Accession format:

- WP_ + 9 digits + version number. For example, WP_000000001.1
- The version number is always '1', WP records will not be updated to a new version but they may be discontinued if no longer found on any genome.

Features of this record type include:

- There is a stable 1:1 correspondence between the accession number and the amino acid sequence.
- Each WP_ record logically represents a protein that is found in one or many genomes from one or multiple species.
- WP proteins will be suppressed if they are no longer annotated on any genome or no longer referenced by any NP or YP record. This processing is not yet implemented.
- If the genome sequence or CDS feature annotation is updated in a manner that changes the protein sequence, then that annotation will point to a new WP_ accession number.
- WP_ records will include protein product name information when available.
- WP_ records include RefSeq as a keyword.
- WP_ records will include Region and Site feature annotation when available from NCBI's Conserved Domain Database.
- WP_ records will include a standard comment:
    - REFSEQ: This record represents a single, non-redundant, protein sequence which may be annotated on many different RefSeq genomes from the same, or different, species.
- WP_ records will not include the 'DBSOURCE' line in GenPept presentation format.
- The initial set of WP_ records will include Source feature annotation of a single species-level tax_id. However, a small sub-set of proteins will represent sequences for more than one species. In the future, these records will be updated to include annotation information for the multiple organisms in which the protein has been identified.
- WP_ records will not include information about the corresponding Nucleotide sequences on the sequence record.
- WP_ records will have links to Nucleotide in the Related Information section of the page display. Links in this section are available through NCBI's E-utilities API.
- Supplemental FTP files will be provided on the RefSeq FTP site reporting protein, nucleotide, and taxonomy data associations. We hope to be able to provide these files with the July RefSeq release or soon thereafter.

Timing:

This new RefSeq product will be included in the July 2013 RefSeq FTP release 60 for WGS bacterial genomes. Previously, WGS bacterial genomes included proteins annotated using the ZP accession prefix. These records are now suppressed and the WGS genome annotation has been updated to refer to the new WP protein products.

RefSeq release 60 will include WP_ protein data in the complete and microbial nodes. These records will be provided using a new file name format and will not be logically bundled with the genomic sequences upon which they are annotated. The file name format will be: microbial.nonredundant_protein.#.{file format type}.gz

Bacterial RefSeq processing is currently beginning to use of the new WP records for all bacterial genomes that represent variation. Thus there will be a transition period during which other RefSeq protein accessions will be suppressed and the corresponding genome annotation will be updated to reference WP protein accessions. This transition is coupled with updating RefSeq annotation using NCBI's RefSeq genome annotation pipeline. We have not finalized the timing of this step; it may occur before RefSeq release 60 or 61.

As noted above, NCBI will continue to provide separate protein records for bacterial RefSeq genomes that are considered to be the best representative(s) for the species; these records will be updated over time to cross-reference the corresponding protein WP accessions but will continue to be provided as a taxonomically organized reference dataset. This new data model is expected to be applied to other taxonomic branches of RefSeq over time.

Bacterial RefSeq Updates – General Timeline:

- Phase 0: implemented; installed with July 2013 RefSeq release 60
    - RefSeq WGS genomes utilizing the NZ_* accession prefix have been updated. Annotated CDS features now refer to WP_ protein accessions. The original protein accessions with the prefix of 'ZP_' have been suppressed.
    - The 'ZP_' accession prefix is deprecated and will no longer be used.
    - Example: Suppressed accession ZP_06483007.1 was previously annotated on NZ_GG700426.1 which was updated to refer to non-redundant protein WP_005733990.1.
- Phase 1: in progress; installed with July 2013 RefSeq release 60
    - NP and YP proteins annotated on RefSeq complete bacterial genomes are being updated to cross-reference WP records. Thus, bacterial RefSeq protein records will adopt the practice of providing a CONTIG line in the sequence block.
- Phase 2: further reduction in redundancy; installed in RefSeq release 61 and/or 62
    - RefSeq genomes will be stratified into two primary categories, that of reference or representative genomes, and those that represent variant genomes that are most redundant and, for instance, generated through population or clonal sequencing approaches.
    - One representative genome will be provided per species. For some species, more than one representative genome will be provided in order to specifically represent a pathogenic strain, or a community standard, or a genome with specific experimental support of the annotation.
    - Representative genomes will continue to cross-reference the NP or YP accession prefix in CDS feature annotation.
    - The additional variant genomes will refer directly to the WP protein identifier.
- Phase 3: re-annotation of most RefSeq bacterial genomes; ongoing processing; will first be included for some genomes in RefSeq release 61 or 62
    - NCBI plans to re-annotate most complete and draft WGS genomes in the RefSeq bacterial genome annotation pipeline. Exceptions will be made for genomes for which annotation is provided through collaboration.
    - Re-annotation using a single framework will result in overall greater consistency in annotation across the RefSeq bacterial genomes dataset.
    - This phase is expected to result in some additions and removals in the WP protein dataset.

Examples:

1. A WP protein may be annotated on multiple genomes.

WP_000289090.1 represents a protein sequence that is identified in, and annotated on, over 500 genomes of Escherichia and Shigella. Each of these genomes has a CDS feature which cross-references this protein accession.

Annotation on NZ_ANMC01000219.1 (Escherichia coli 97.0007):

```
CDS             complement(10258..10452)
                /gene="bfd"
                /locus_tag="EC970007_3838"
                /codon_start=1
                /transl_table=11
                /product="bacterioferritin-associated ferredoxin"
                /protein_id="WP_000289090.1"
                /db_xref="GI:446211235"
                /translation="MYVCLCNGVSDKKIRQAVRQFHPQSFQQLRKFIPVGNQCGKCIR
                AAREVMQDELMQMPEFKEIA"
```

Annotation on NZ_ADUT01000021.1 (Shigella dysenteriae):

```
CDS             complement(452012..452206)
                /locus_tag="SD1617_2124"
                /codon_start=1
                /transl_table=11
                /product="bacterioferritin-associated ferredoxin"
                /protein_id="WP_000289090.1"
                /db_xref="GI:446211235"
                /translation="MYVCLCNGVSDKKIRQAVRQFHPQSFQQLRKFIPVGNQCGKCIR
                AAREVMQDELMQMPEFKEIA"
```

2. Interim processing may add then remove redundant WP proteins – and - partial CDS features at the end of a contig may refer to a longer WP

During the transition phase to shift to the new WP data model, some records may be updated more than once. For example, an genome may be updated to remove references to ZP proteins (which were then suppressed) and add references to WP proteins and updated again to change the WP references for partial CDS features that are at the end of the contig without changing the annotated CDS coordinates. A similar type of update (to change the WP reference) will occur if the CDS coordinates, or nucleotide sequence within the CDS region, are updated over time (resulting in a different protein sequence). This example illustrates the former, which occurs during the initial phase of moving to this new data model but is not expected to continue once the transition is complete.

The CDS annotation on NZ_AJGF01000004.1 (Myocbacterium abscessus M93) was updated on May 8, 2013 to remove references to ZP accessioned proteins (which were suppressed) and to add references to WP proteins. NCBI's revision history display can be used to compare and view older versions of this genomic record.

Focusing on a single gene with locus_tag "OUW_04031", the May 8th update replaced ZP_14236142.1 with WP_005080260.1 , a 398 aa protein; the original ZP protein and the replacement WP protein are identical in sequence and length. This update step reflects the initial phase of the process, converting RefSeq WGS genomes to the WP data model.

```
     CDS             224972..>226166
                     /locus_tag="OUW_04031"
                     /note="COG1290 Cytochrome b subunit of the bc complex"
                     /codon_start=1
                     /transl_table=11
                     /product="ubiquinol-cytochrome c reductase cytochrome b
                     subunit"
                     /protein_id="ZP_14236142.1"
                     /db_xref="GI:419708673"
```

```
     CDS             224972..>226166
                     /locus_tag="OUW_04031"
                     /note="COG1290 Cytochrome b subunit of the bc complex"
                     /codon_start=1
                     /transl_table=11
                     /product="ubiquinol-cytochrome c reductase cytochrome b
                     subunit"
                     /protein_id="WP_005080260.1"
                     /db_xref="GI:491221950"
```

Similar processing of RefSeq WGS genome NZ_AKTX01000006.1 (Mycobacterium abscessus 5S-0304) to remove ZP accessions and replace with the new data model WP accessions, resulted in replacing ZP_15341076.1 with WP_005099714.1, a 543 aa protein which is identical to, but longer than, the above protein WP_00508260.1.

Post-analysis to identify redundant WP proteins that are identical in sequence but differ in length specifically due to unavailable nucleotide sequence data (the end of a contig) resulted in a second update to NZ_AJGF01000004.1 (Myocbacterium abscessus M93) on June 3 2013 to remove the reference to partial protein WP_00508260.1 and replace it with a reference to the complete protein WP_005099714.1 (image below). Future processing will identify WP proteins that are no longer annotated on any genome (nor referred to by any NP or YP accession) and suppress them. Thus, WP_005099714.1 is annotated on two related genomes; in one case as a complete CDS (NZ_AKTX01000006.1) because the contig sequence completely represents the CDS region, and in the other as a partial CDS (NZ_AJGF01000004.1) at the end of the contig sequence (note the CDS coordinates and contig coordinates marked with arrows below).

```
     CDS             224972..>226166
                     /locus_tag="OUW_04031"
                     /EC_number="1.10.2.2"
                     /note="COG1290 Cytochrome b subunit of the bc complex"
                     /codon_start=1
                     /transl_table=11
                     /product="menaquinol-cytochrome C reductase cytochrome b
                     subunit"
                     /protein_id="WP_005099714.1"
                     /db_xref="GI:491241505"
                     /translation="MSDTAQKPSRAAKQAEAMDSRYHLAAGMKRQINKVFPTHWSFML
                     GEIALYSFIVLLLSGVYLTLFFDPSMSEVTYNGIYQPLRGVQMSKAYETTLNISFEVR
                     GGLFVRQIHHWAALMFAASIMVHMARIFFTGAFRRPREANWVIGALLFILAMFEGFFG
                     YSLPDDLLSGTGIRAALSGITMGLPLIGTWMHWALFGGDFPGNILIPRLYAMHILLIP
                     AIILALIGIHLALVWYQKHTQFPGPGATEKNVVGVRILPVFALKGGSFFAFTTAILAL
                     MSGLLQINPIWVLGPYKPSQISAGSQPDFYMMWTDGLLRIIPAWEIYPFGHTIPQAVW
                     VAVGMGLVFGLLIAYPFLEKKLTGDDAHHNLLQRPRDAPVRTAIGSAAISLYMLFTLM
                     CMNDII"
CONTIG       join(AJGF01000004.1:1..226166)
```

3. A NP protein record for a representative genome refers to a WP protein.

Proteins annotated on the *Escherichia coli K12* reference genome will continue to be tracked with existing NP accessions which in turn refers to the non-redundant WP record. The E. coli K12 annotated genome record, NC_000913.2, includes CDS feature annotation for transaldolase:

```
CDS             8238..9191
                /gene="talB"
                /locus_tag="b0008"
                /gene_synonym="ECK0008; JW0007; yaaK"
                /EC_number="2.2.1.2"
                /function="enzyme; Central intermediary metabolism:
                Non-oxidative branch, pentose pathway"
                /function="1.7.3 metabolism; central intermediary
                metabolism; pentose phosphate shunt, non-oxidative branch"
                /function="7.1 location of gene products; cytoplasm"
                /GO_component="GO:0005737 - cytoplasm"
                /GO_process="GO:0009052 - pentose-phosphate shunt,
                non-oxidative branch"
                /codon_start=1
                /transl_table=11
                /product="transaldolase B"
                /protein_id="NP_414549.1"
```

NP_414549.1 includes rich feature annotation including and refers to the non-redundant protein record WP_001264707.1 in the Protein feature and the sequence block (the CONTIG line). In the near future, we will present these records with both the CONTIG line and the ORIGIN line with sequence.

```
     Protein         WP_000130185.1:1..317
                     /product="transaldolase"
                     /calculated_mol_wt=35088
CONTIG       join(WP_000130185.1:1..317)
```

WP_001264707.1 is also annotated on 59 *Escherishia coli* genomes including NZ_ANXP01000003.1.

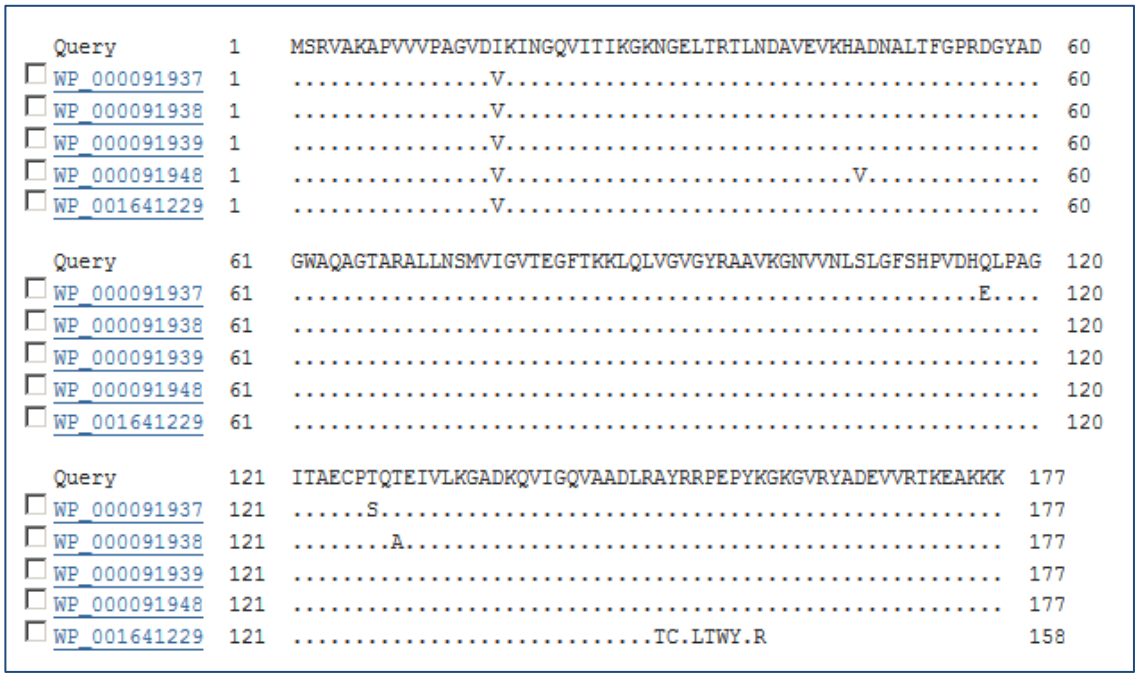
4. A cluster of WP accessions may be logically related but represent distinct sequences.

The following five WP accessions represent five specific sequences of the 50S ribosomal protein found in Salmonella. These sequences differ by individual amino acids or length and in the number of genomes that they have been annotated on. These sequences (and more) are found in the same protein cluster and are logically related at the level of sequence; proteins in the same cluster also likely have the same function. Currently, individual WP records do not include feature annotation or links to the protein cluster but we anticipate providing this information in the near future.

- WP_000091935.1
  - 177 aa, 50S ribosomal protein L6 [Salmonella enterica]
  - Annotated on NZ_CM001471.1

- WP_000091937.1
  - 177 aa, 50S ribosomal protein L6 [Salmonella bongori]
  - Identical to reference protein YP_004731847.1 which is annotated on the representative genome for this species (NC_015761).
- WP_000091938.1
  - 177 aa, 50S ribosomal protein L6 [Salmonella enterica]
  - Annotated on more than one genome including NZ_KB731367.1, NZ_CM001151.1
- WP_000091939.1
  - 177 aa, 50S ribosomal protein L6 [Salmonella enterica]
  - Annotated on over 300 Salmonella genomes including NZ_CATS01000166.1
- WP_000091948.1
  - 177 aa, 50S ribosomal protein L6 [Salmonella enterica]
  - Annotated on over 600 genomes from Escherichia coli (NZ_KB733094.1) and Shigella flexneri (NZ_AFHB01000042.1)
  - Identical to reference protein YP_001573135.1 which is annotated on a Salmonella enterica genome (NC_010067.1)
- WP_001641229.1
  - 165 aa, LSU ribosomal protein L6p [Salmonella enterica]
  - Annotated on NZ_AFCS01001059.1

Each of these 5 WP accessions represents a unique, but related, sequence. Sequence differences are seen in a Neeleman-Wunsch protein multiple alignment, using WP_000091935.1 as the query:



```
Query          1    MSRVAKAPVVVPAGVDIKINGQVITIKGKNGELTRTLNDAVEVKHADNALTFGPRDGYAD  60
WP_000091937   1    .................V.........................................  60
WP_000091938   1    .................V.........................................  60
WP_000091939   1    .................V.........................................  60
WP_000091948   1    .................V............................V.............  60
WP_001641229   1    .................V.........................................  60

Query          61   GWAQAGTARALLNSMVIGVTEGFTKKLQLVGVGYRAAVKGNVVNLSLGFSHPVDHQLPAG  120
WP_000091937   61   ...........................................................E....  120
WP_000091938   61   ............................................................  120
WP_000091939   61   ............................................................  120
WP_000091948   61   ............................................................  120
WP_001641229   61   ............................................................  120

Query          121  ITAECPTQTEIVLKGADKQVIGQVAADLRAYRRPEPYKGKGVRYADEVVRTKEAKKK  177
WP_000091937   121  ......S.....................................................  177
WP_000091938   121  ........A...................................................  177
WP_000091939   121  ............................................................  177
WP_000091948   121  ............................................................  177
WP_001641229   121  ............................TC.LTWY.R                        158
```