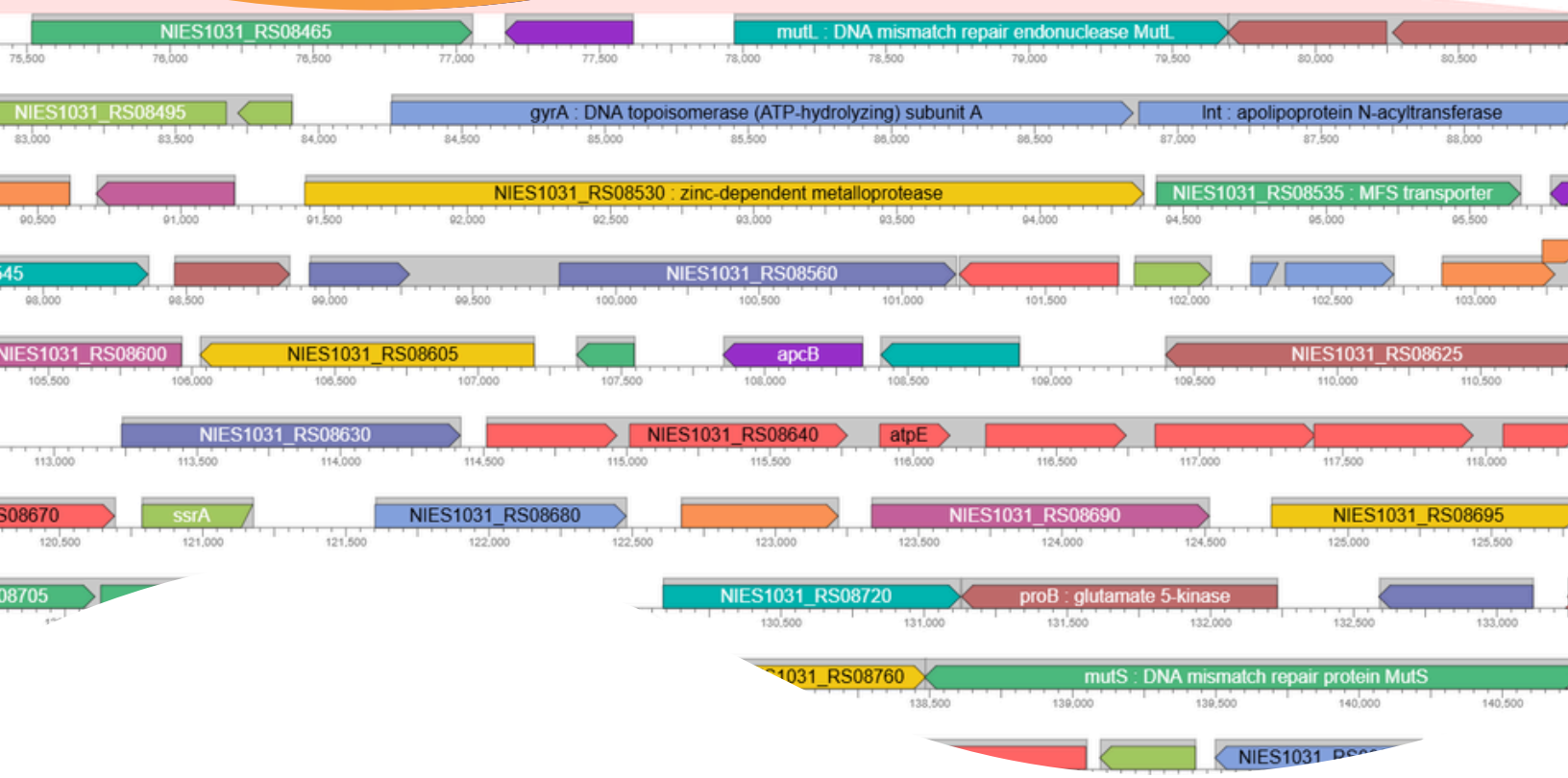
BioCyc



Manual Curation in BioCyc

Don’t Let Database Errors Compromise Your Next Paper

Better Science, Faster

Manual Curation in BioCyc | Don't Let Database Errors Compromise Your Next Paper

Caspi R., Wilson-Mortier B., Beesley D., Karp P.

Abstract

BioCyc databases integrate a diverse array of information by combining computationally inferred information, data imported from other databases, and manual curation. We quantify the amount of manual curation in BioCyc by comparing protein annotations at the start and end of the curation process. Across 10 moderately curated BioCyc databases, 7,126 protein annotations were modified by curators, or 27.1% of the 26,248 proteins encoded in the 10 genomes. Although the existence of annotation errors in the public genome databases is well known, we expect that most researchers would be surprised that the number of errors is this high. BioCyc curation therefore provides a substantial elevation in quality beyond the public genome databases. This reduction in errors also yields a significant reduction in the financial and experimental risk of research relying on uncurated genome databases.

Introduction to BioCyc Database Construction

The BioCyc.org genome and metabolic pathway web portal creates exceptionally rich and accurate portraits of sequenced organisms by combining computational inferences, data imports, and manual curation. These organism-centered databases enable fuller understanding of gene functions by embedding genes in their pathway and regulatory contexts. BioCyc provides an evidence-based foundation of information distilled from more than 167,000 publications that enable more accurate hypothesis formation, improved interpretation of experiments, and higher quality grant proposals and publications.

BioCyc database creation begins with the annotated genomes provided by NCBI RefSeq. Our extensive bioinformatic pipeline supplements this data with:

- Predicted transport reactions, metabolic reactions, and metabolic pathways
- Inferred operons
- Inferred protein complexes
- Candidate genes encoding missing metabolic enzymes (pathway hole filler)
- Cross-genome protein ortholog group assignments.

Data import tools integrate information from other bioinformatics databases and publications, including available Gene Ontology annotations and protein feature data from UniProt, gene essentiality datasets, transcription start sites, ribosome binding sites, and terminators.

Database Curation in BioCyc

Even for well-studied model organisms, automated genome annotation pipelines such as RefSeq's Prokaryotic Genome Annotation Pipeline (PGAP), are prone to misannotations. As a result, proteins that have undergone substantial experimental characterization are often labeled as "hypothetical proteins" or assigned incorrect functions in uncurated databases. When an upstream annotation is incorrect, that error can cascade across thousands of downstream records. Misannotated genes can lead to incorrect pathway reconstruction, flawed metabolic models, misdirected strain engineering strategies, and wasted reagent, instrument, and personnel time.

To combat this problem, BioCyc applies extensive manual curation to improve the quality and accuracy of selected high-priority databases. This task results in substantial and measurable changes, often involving hundreds or thousands of updates per database. These updates include, but are not limited to, the addition, removal, or modification of database objects such as gene and protein annotations, enzymatic reaction assignments, and predicted metabolic pathways.

The gold standard of evidence for such modifications in BioCyc is experimental data. Accordingly, manual curation requires domain expertise and

Figure 1 Genome browser figure from B. subtilis 168 showing how this imported info is all shown together for several adjacent operons

comprehensive review of the experimental literature. In some cases, curators also rely on strong computational evidence to infer the functions of hypothetical or misannotated proteins.

To quantify the impact of curation, we performed detailed comparative analyses of 10 databases before and after curation. This analysis quantifies:

1. Changes in protein names (functions)
2. Changes in reaction assignments for metabolic enzymes
3. Changes in the presence or absence of metabolic pathways

Results

The results of these comparisons are shown in **Table 1**. For example, the genome of *Lactiplantibacillus plantarum* WCFS1, an industrially significant and highly characterized lactic acid bacterium (LAB), received 958 updated protein annotations (38.7% of the proteome), 69 pathway additions and 74 removals (143 total pathway modifications; 69.4% of pathways present before manual curation), and 557 modifications to metabolic enzyme reactions (57.3% of all metabolic enzymes). In total, 1,658 modifications were made to this database, demonstrating that even a well-characterized strain can benefit from manual curation.

Across all 10 genomes in **Table 1**, BioCyc curators updated 7,126 protein annotations, representing 27.1% of the aggregate proteome (26,248 proteins).

Although many of the alterations to protein names do reflect changes in the functions assigned to these proteins, some changes reflect stylistic concerns, such as to conform with BioCyc protein

Table 1: Analysis of curation changes performed during manual curation of 10 BioCyc databases.

Organism	Genome Size (proteins)	Protein Annotation Changes	Percent of Proteins Modified	Total Enzymes	Enzymes Altered	Percent of Enzymes Altered	Pathways At Start	Pathways Added or Removed
***Acinetobacter baumannii* AB5075**	3639	733	20.1	1083	468	43.2	242	63
***Chlamydia trachomatis* D/UW-3/CX**	889	110	12.4	297	74	24.9	80	20
***Enterococcus faecalis* V583**	3201	742	23.2	882	393	44.6	237	109
***Helicobacter pylori* 26695**	1467	524	35.7	666	379	56.9	162	79
***Lactiplantibacillus plantarum* WCFS1**	3024	958	31.7	972	557	57.3	206	143
***Neisseria gonorrhoeae* FA 1090**	1940	496	25.6	766	306	39.9	188	76
***Staphylococcus aureus* NCTC 8325**	2836	1105	39	858	479	55.8	153	91
***Streptococcus pneumoniae* D39V**	1993	608	30.5	769	316	41.1	236	107
***Synechocystis* sp. PCC 6803**	3691	883	23.9	1108	649	58.6	243	124
***Vibrio cholerae* O1 biovar El Tor str. N16961**	3568	967	27.1	1221	502	41.1	303	84

naming conventions. We reviewed 2,000 proteins whose names were changed across 10 curated BioCyc databases, and found that an average of 50.03% of the changes were significant functional changes, as opposed to stylistic changes. Therefore, we infer that approximately 50.03% of the 7,126 proteins modified across the 10 genomes received significant changes to their functions, namely 3,565 (13.6% of the aggregate proteomes), or 356 proteins per genome Examples of the types of annotation changes made to *L. plantarum* WCFS1, both functional and stylistic, can be seen in **Table 2.**

Significantly, the comparisons shown in **Table 1** and **2** also underestimate the full value added by manual curation. In addition to the changes quantified above, curators also incorporate experimentally derived data, including:

1. Gene essentiality (under defined conditions)
2. Regulatory information (transcription factor binding sites and regulons)
3. Expression data from transcriptomic and/or proteomic datasets
4. Small regulatory RNA datasets
5. Novel and previously unannotated genes and proteins

Conclusions

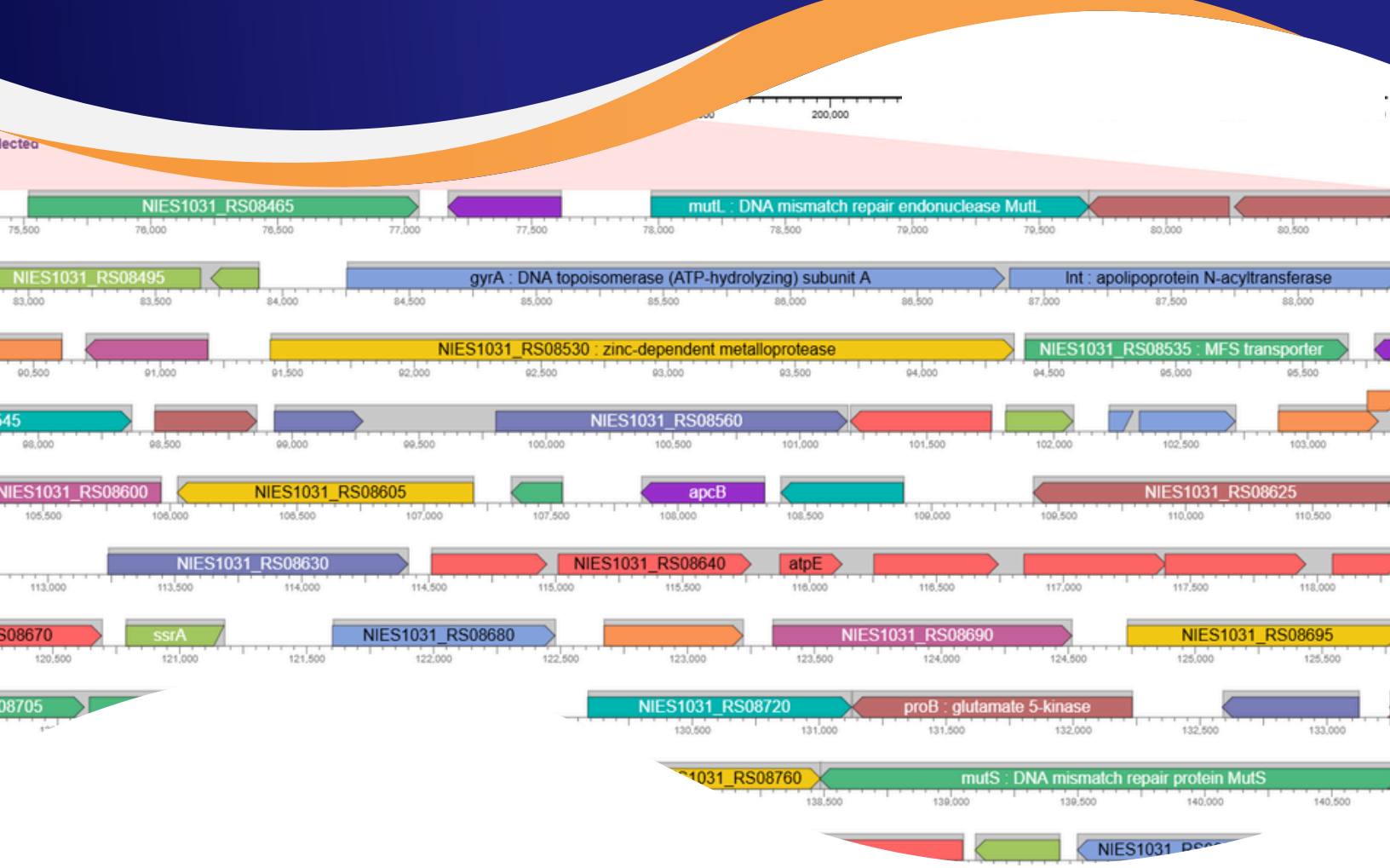
These curation practices elevate the quality of BioCyc's manually curated databases well beyond that of standard genome databases. In addition, curation provides original, high-quality syntheses of current knowledge on well-characterized proteins and pathways, including references to relevant primary literature. Without expert curation and literature validation, the probability that annotation errors will undermine experimental results, publications, and grant proposals (with both financial costs and loss of productivity) increases substantially.

Table 2: Examples of protein annotation updates in the genome of *L. plantarum* WCFS1

Accession	Initial Function	Revised Function	Type of Change	Justification
lp_1229	MucBP domain-containing protein	mannose-specific adhesin	Assignment of function	Function described in literature (PMID 16109954, PMID 21840797)
lp_1204	EpsG family protein	capsular polysaccharide 2 polymerase Cps2H	Assignment of function	Function described in literature (PMID 23170998)
lp_0185	PTS beta-glucoside transporter subunit IIBCA	Sucrose-specific PTS 1 enzyme IIBCA	Change of function	Function described in literature (PMID 17261521)
lp_3479	LacI family DNA-binding transcriptional regulatorr	Galactose metabolism transcriptional regulator GalR	Change of generic function to specific function	Function described in literature (PMID 32059863)
lp_2514	MFS transporter	Nicotinamide transporter	Change of generic function to specific function	Function described in literature (PMID 40727558)
lp_2370	F0F1 ATP synthase subunit A	F1Fo-ATPase subunit a	Change of name format	The enzyme acts in lactic acid bacteria primarily as an ATP-dependent proton pump (ATPase) rather than an ATP synthase (PMID 10940012). In addition, the correct format for the subunits of the Fo component is lower case letters
lp_0796	Alpha/beta fold hydrolase	Hydroxycinnamic acid esterase	Assignment of function	Function described in literature (PMID 23793626, PMID 23054632)

Table 2: Examples of protein annotation updates in the genome of *L. plantarum* WCFS1 *continued*

Accession	Initial Function	Revised Function	Type of Change	Justification
lp_0105	Nickel pincer cofactor biosynthesis protein LarB	Pyridinium-3,5-biscarboxylic acid mononucleotide synthase	Change of generic function to specific function	Function described in literature (PMID 27114550)
lp_0585	sigma-54-dependent transcriptional regulator	sigma54 activator protein	change of name format	the annotated name was misleading, as the protein is not dependent on sigma-54, but rather activates it (PMID 19942662)
lp_0418	Response regulator transcription factor	Plantaricin A-sensing response regulator PlnD	Change of generic function to specific function	Function described in literature (PMID 11532131, PMID 12519198, PMID 19519894)
lp_1814	L-lactate permease	Lactate:proton symporter	Change of name format with additional information	The enzyme is very similar (E-value of 6e-30) to the *E. coli* enzyme, which is known to be a symporter (PMID 11785976)
lp_2919	Proline iminopeptidase-family hydrolase	Prolinase 2 monomer	Change of name format	Name was changed to match the literature. “monomer” was added to distinguish from the dimeric complex (PMID 25463046)
lp_0268	CDP-glycerol glycerophosphotransferase family protein	Teichoic acid poly(glycerol phosphate) polymerase 1	Change of name format	Name modified to correspond to EC name. Serial number added since the organism has two isofroms (PMID 23454138).



BioCyc

Genome Database Collection

Better Science, Faster