AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere

Check for updates

| Editor's Pick | Genomics and Proteomics | Research Article

Differences in GenBank and RefSeq annotations may affect genomics data interpretation for *Pseudomonas putida* KT2440

Guilherme Marcelino Viana de Siqueira,[1,2,3] Thomas Eng,[2,3] Aindrila Mukhopadhyay,[2,3,4] María-Eugenia Guazzaroni[1]

AUTHOR AFFILIATIONS See affiliation list on p. 9.

ABSTRACT Annotations of genomic features are cornerstone data that support routine workflows in conventional omics analyses in *Pseudomonas putida* KT2440 and other organisms. The GenBank and the RefSeq versions of the annotated KT2440 genome are two popular resources widely cited in the literature; yet, they originate from distinct prediction pipelines and possess potentially different biological information that is often overlooked. In this study, we systematically compared the features present in these resources and found that approximately 16% of the total of KT2440 open reading frames (ORFs) show differences in their predicted genomic positions across GenBank and RefSeq, despite sharing equivalent locus tag codes. Furthermore, we show that these discrepancies can affect the results of high-throughput analyses by processing a collection of RNAseq expression data sets utilizing both annotations. Our findings provide a comprehensive overview of the current state of available resources for genomics research in *P. putida* KT2440 and highlight a rarely addressed yet widespread potential pitfall in the literature on this organism, with possible implications for other prokaryotes.

IMPORTANCE Genome annotation databases often rely on different statistical models for their function predictions and inherently carry biases propagated into studies using them. This work provides a quantitative assessment of two popular annotation resources for the model bacterium *Pseudomonas putida* KT2440 and their influence on data interpretation. As large-scale omics data sets are commonly used to inform experimental decisions, our results aim to promote awareness of the caveats associated with these computational resources and foster reproducibility and transparency in *P. putida* research.

KEYWORDS *Pseudomonas putida*, RefSeq, genome annotations, GenBank

Pseudomonas putida KT2440 is a model saprophytic Pseudomonad strain that has been widely studied due to its innate metabolic versatility, including the ability to degrade a vast array of chemicals, ease of genetic manipulation, and its latent safety as a non-pathogenic microbe (1–3). The first version of the *P. putida* KT2440 reference genome (accession AE015451), published in the GenBank database by Nelson et al. in the early 2000s (4), greatly contributed to this popularity and rapidly accelerated research on this strain, enabling the construction of several metabolic models (5–9), and the leveraging of *P. putida*'s genetic traits for bioengineering efforts (1).

After the publication of the original genome sequence, the National Center for Biotechnology Information (NCBI) incorporated this assembly in its own reference sequence (RefSeq) database under the accession NC_002947. The RefSeq database is an NCBI project that offers a curated set of high-quality, non-redundant reference records, combining chromosomal, transcript, and protein information, as well as structural

Editor Grant R. Bowman, University of Wyoming, Laramie, Wyoming, USA

Address correspondence to María-Eugenia Guazzaroni, meguazzaroni@ffclrp.usp.br.

The authors declare no conflict of interest.

See the funding table on p. 10.

Received 4 June 2025
Accepted 4 September 2025
Published 2 October 2025

and functional information and other metadata, for viruses, microbes, organelles, and eukaryotic organisms (10).

In 2021, RefSeq included over 200,000 bacterial and archaeal genomes (11). To ensure the quality of these annotations, it relies on a unified pipeline, the "Prokaryotic Genome Annotation Pipeline" (PGAP), which combines homology-based and *ab initio* methods for predicting functional elements directly from sequences (12). As a means to standardize nomenclature and reduce redundancies across the entire database, in 2015, the RefSeq team took on the massive effort of reannotating all of the prokaryotic genome assemblies deposited in the database using PGAP version 3 (13). This was done again in 2017 following the release of PGAP version 4.1 (13). Currently, the RefSeq team adopts a rolling schedule in which the oldest live assemblies are automatically re-annotated daily to benefit from updates incorporated into the pipeline (11).

Unlike RefSeq records, GenBank records are not revisited as often. In 2016, the GenBank record of the KT2440 genome annotation received its only major update to date (accession AE015451.2). This version is based on a resequencing of the original 2002 strain, with the employment of different computational tools (AMIGene and Prodigal), RNAseq expression data, and the manual curation of over 1,500 genes for the structural re-annotation of the genome (14). The new annotation includes over 300 newly predicted genes compared to the original record and lists 5,786 features, the vast majority of which (5,564) are protein-coding genes, identified by genomic loci codes with a "PP_" prefix. In contrast, the latest RefSeq annotation for *P. putida* at the time of writing this manuscript (accession NC_002947.4), from December 2023, accounts for 5,693 annotated features, 5,486 of which are protein-coding genes that are named using a RefSeq "PP_RS" locus tag re-annotation prefix.

RefSeq is generally regarded as a reliable annotation source for microbial genomics (11), and the RefSeq annotation of the *P. putida* KT2440 genome is frequently cited in the literature. However, GenBank's locus tag prefix format is often preferred to refer to specific genes in peer-reviewed publications—even when authors use recent RefSeq annotations—and the curated information provided by RefSeq appears to be unevenly propagated to other popular knowledge bases, including PseudomonasDB (15), BioCyc (16), Uniprot (17), Kyoto Encyclopedia of Genes and Genomes (KEGG) (18), iModulons (19, 20), and the Fitness Browser (21). Considering the way both annotation sources are used with the presumption that they are truly interchangeable in the *P. putida* KT2440 literature, here we systematically compare the features present in each of them to assess the possible impacts of blindly choosing either for the interpretation of omics data sets in works utilizing *P. putida* KT2440 as a model strain.

MATERIALS AND METHODS

Cross-referencing *P. putida* KT2440 genome annotations

P. putida KT2440 RefSeq and GenBank records were downloaded from the NCBI's FTP server in September 2024 from assemblies GCF_000007565.2 and GCA_000007565.2, respectively. The total number of protein-coding genes in each annotation, as well as information on their genomic positions and other features of interest, was extracted from the *_feature_table.txt.gz file available in each directory. The resulting tables were processed in R (version 4.4.2) using the Tidyverse suite of packages (version 2.0.0) (22). The retrieval of data from KEGG to compose Table 1 was made using the KEGGREST R package (version 1.44.1) (23). The functional categorization of position-shifted genes was performed using annotations from the cluster of orthologous genes (COG) database (24). The treemap visualization for these functional annotations was made using the package treemap (version 2.4.4) (25) in R. The overrepresentation analysis for COG annotations was conducted using a one-tailed Fisher's exact test. All of the data processing steps and the code used in this analysis are available on GitHub (https://github.com/GuazzaroniLab/ppuGenomeAnnotations).

TABLE 1 Overview of major online genome resources for *P. putida* KT2440

Database	Associated annotations	Notes	Website
ALEdb	AE015451; NC_002947	Includes other annotations of KT2440 genomes, depending on the experiment source	https://aledb.org/
BioCyc	NC_002947	A 2016 version of the RefSeq annotation was used to populate the *P. putida* database, created in 2017	https://biocyc.org/
Fitness Browser	AE015451	Provides their own set of curated re-annotations for over 40 *P. putida* genes based on new experimental data	https://fit.genomics.lbl.gov/
KEGG	AE015451	Out of 5,786 entries (genes) on KEGG, only 1,804 are associated with metabolic pathways	https://www.kegg.jp/
iModulonDB	AE015451	Categorized the genome of *P. putida* KT2440 into 84 groups of independently modulated genes utilizing a machine-learning approach	https://imodulondb.org/
UniProt	AE015451; NC_002947	Offers hyperlinks to both GenBank and RefSeq for any given accession, although RefSeq locus tags themselves are not recognized as search terms	https://www.uniprot.org/

Collecting historical data on genome annotations usage in the literature

Citation trends for the RefSeq and GenBank *P. putida* KT2440 genome annotations were obtained from Google Scholar search results using the terms "NC_002947" and "AE015451," respectively. The metadata from all of the indexed publications was captured using a custom R script using RSelenium (version 1.7.9) (26). The resulting database was curated to remove duplicate entries, preprints, and other low-quality results. The final table is available as a plain text file (File S1). The code used to retrieve and process publication metadata is available on GitHub (https://github.com/GuazzaroniLab/ppuLiteratureMetrics).

Transcriptomic data processing

To compare the impacts of using either annotation source in the calculated results of a standard RNAseq pipeline, we downloaded data from several different publicly available RNAseq data sets from NCBI's Sequence Read Archive (19, 27–33) (Table S1), composing a collection of different biological conditions and sequencing library layouts, and processed them using both the RefSeq and the GenBank versions of the transcriptome as reference.

In short, for this analysis, raw sequencing data (FASTQ format) were downloaded using the NCBI fasterq-dump command line utility (version 2.11.3). The retrieved data were then further processed for the removal of sequencing adapters and low-quality base assignments with Trimmomatic (version 0.39) (34). RNA-Seq by expectation maximization (RSEM) (35) was used to map reads to either the RefSeq or the GenBank version of the reference *P. putida* KT2440 transcriptome, and the DESeq2 package (version 1.44.0) (36) was used for the determination of differentially expressed genes.

The code used for retrieving and processing the expression data sets for this analysis is available on GitHub (https://github.com/GuazzaroniLab/ppuRNAseqCollection). The resulting expression data sets (GenBank- and RefSeq-processed) are available as plain text files in this repository.

RESULTS

Both GenBank and RefSeq genome annotations are widely used in the literature

To verify historical trends in the usage of the GenBank or RefSeq *P. putida* KT2440 genome accession codes in the literature, we retrieved Google Scholar results for the queries "AE015451" and "NC_002947," respectively. As of December 2024, out of the roughly 400 selected publications retrieved after curation, 178 (43%) included the GenBank *P. putida* KT2440 genome accession code, whereas 238 (57%) publications included the RefSeq identifier. Historical data show that RefSeq has always been more predominant than GenBank in the number of citations by a small margin (Fig. 1A), but this seems to have become more prominent around 2016, coinciding with the release

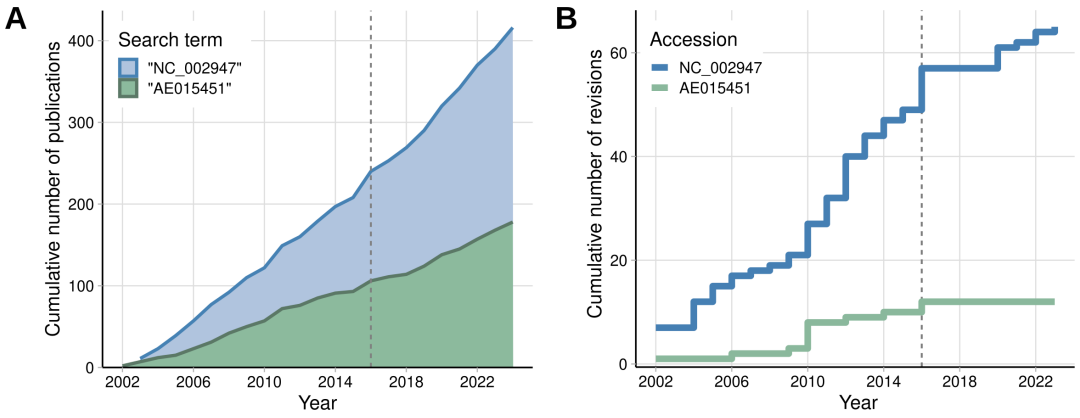

FIG 1 Historical trends for citations and the version history of GenBank and RefSeq *P. putida* KT2440 genome annotations. (A) Google Scholar search results for publications containing the search terms "NC_002947" (RefSeq) or "AE015451" (GenBank) over time. (B) Revision history of the RefSeq (blue) and GenBank (green) genome annotation records in the Nucleotide NCBI database (nuccore). In both panels, the year 2016, when both NC_002947.4 and AE015451.2 were introduced, is indicated by the vertical dashed line.

of the RefSeq annotation version 4 (NC_002947.4), the current major version as of the time of the writing of this manuscript. In total, the RefSeq version of KT2440's genome annotation has received four major updates since 2002, and a total of 65 version changes in total (Fig. 1B). The GenBank record, as previously noted, is not updated as often and has received only one major version change in 2016 (14), which is still the current version.

KT2440 genes are historically associated with a "PP" or "PP_" prefix followed by a four-digit number (4, 14). Owing to the Prokaryotic RefSeq Genome Re-annotation initiative announced by NCBI in 2015, recent RefSeq annotations introduce a new locus tag code system for their own assigned ORFs using the prefix "PP_RS" followed by a five-digit number (37). Due to the many updates to the entire RefSeq database and the difficulty in retrieving older versions from the website, it is hard to ascertain when the RefSeq version of the *P. putida* KT2440 genome annotation started to include this new locus tag system. However, the PGAP annotation pipeline provides a useful direct correspondence between new RefSeq locus tags and the GenBank annotation tags whenever an equivalent CDS is re-annotated on the new record. For instance, under this new nomenclature, PP_0001 is equivalent to PP_RS00005, PP_0002 to PP_RS00010, etc.

Relying on these conversions can, however, become a source of confusion as the gene products for a given locus tag might receive different nomenclatures across different knowledge bases depending on the underlying source it pulls data from (Table 1). For instance, several hypothetical proteins in GenBank were later reannotated with more specific nomenclature using PGAP, such as PP_1173 ("porin-like protein" in GenBank and "OprD family porin" in RefSeq), and PP_2747 ("conserved protein of unknown function" in GenBank and "AMP-binding protein" in RefSeq). A search for these genes in UniProt will often match either the GenBank (in the case of PP_1173) or RefSeq (in the case of PP_2747) names, whereas in databases like KEGG, they only coincide with the GenBank nomenclature. Moreover, the new RefSeq locus tag codes are also not widely recognized as search terms in any of the mentioned databases. For instance, data from ALEdb, a database that aggregates data from several adaptive laboratory evolution projects (38), show that the vast majority of unique mutations recorded in the database come from genomes analyzed using the RefSeq NC_002947.x family of annotations as a reference (Fig. S1), even though the genes themselves are only searchable in the database with the GenBank code. This favors the maintenance of the GenBank codes as the standard way to refer to specific loci in the *P. putida* KT2440 genome, despite the possibility of converting between old and new locus tag annotations, which could

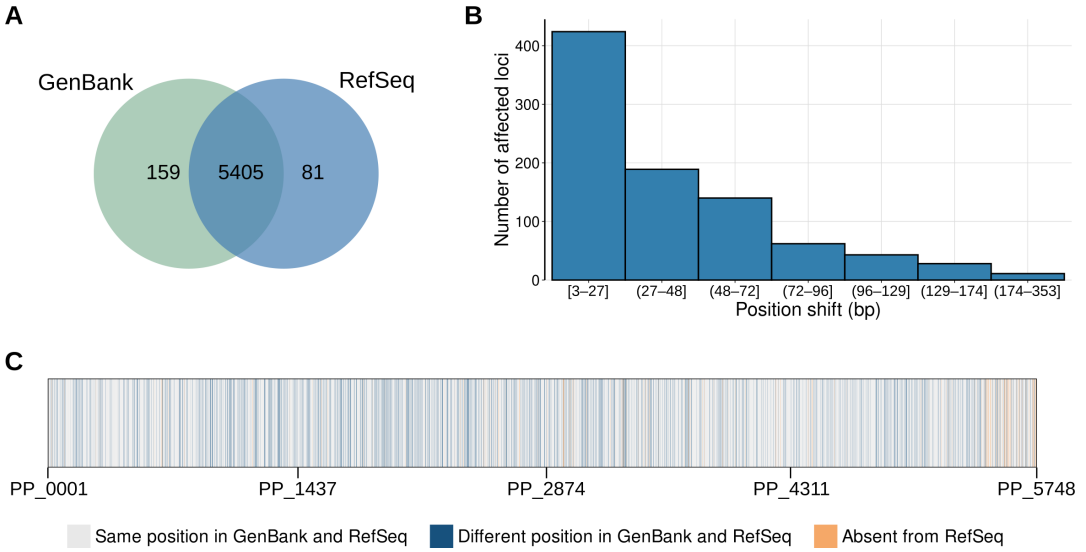

FIG 2 Comparison between features in GenBank and RefSeq *P. putida* KT2440 annotations. (A) Venn diagram comparing the number of features (genes) in GenBank (green) and RefSeq (blue) with the "protein-coding" descriptor. (B) Position differences in the start or end nucleotide of the shifted ORFs (in base pairs). (C) Categorization of genes across all of the possible GenBank loci in comparison to their RefSeq counterpart. For these comparisons, RefSeq re-annotation locus tag codes were matched to the GenBank locus tag codes by using the "old_locus_tag" field whenever available. Please note that panel C does not accurately reflect the genome structure of *P. putida* KT2440, as locus tag codes do not necessarily match their relative genomic positions.

obscure more pressing differences between both sources, as explored in more detail in the next sections.

The genomic coordinates of an expressive number of genes are different in GenBank and RefSeq

Even though it is possible to map annotations between RefSeq and GenBank, we find that both the number of features annotated as "protein-coding" (genes) and their positions can vary greatly between both resources, as seen in Fig. 2A. Importantly, while the vast majority (5,045) of annotated genes are present in both, 897 annotations with equivalent locus tag codes differ either in the start or end genomic positions. Almost half of these position differences comprise 27 base pairs or fewer, although a few cases reach a couple hundred base pairs (Fig. 2B; File S2), for loci distributed all over the genome (Fig. 2C). We found that the re-annotation did not introduce any frameshifts to ORFs, except for the transposases in loci PP_2977 and PP_4420. Yet, as a result, 455 annotated genes had a different start codon position and 442 had a different stop codon position between both annotations, effectively changing the transcript size predicted by each of them, with RefSeq ORFs being smaller in the majority of the cases (Table S2).

There are 159 genes in the GenBank annotation that were later excluded from the RefSeq annotation. On the other hand, RefSeq introduces its own set of 81 genes without a direct counterpart in GenBank. In both cases, most of the genes were annotated as conserved proteins of unknown function and hypothetical proteins. However, unlike GenBank or RefSeq exclusive genes, the set of genes with differences in their positions along the genome (or shifted genes) is much more diverse. It includes enzymes, transporters, and transcriptional regulators among other classes of proteins.

To better understand the diversity of biological functions associated with the shifted genes, we utilized the latest set of COG classifications available in NCBI (24) and compared the classification of genes within the shifted set with those that remain in the same position in both annotations. Among the 897 shifted genes, only 260 did not have COG annotations associated with them (Fig. 3), and Fisher's exact test was used to determine whether a statistically significant bias existed toward any of the five COG functional categories occurring in the subset of genes with associated annotations.

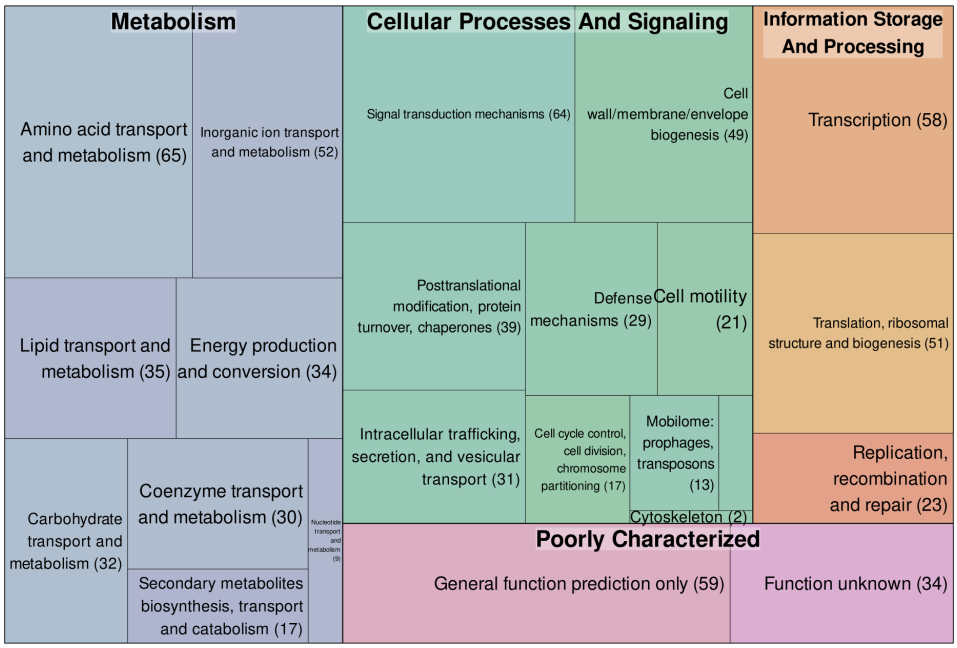

FIG 3 COG classification for all functionally annotated shifted genes. The area of the squares correlates to the total number of genes associated with each category, which is also indicated in parentheses. Smaller boxes with similar colors correspond to COG term descriptions that fit within the same functional categories (indicated in bold).

We found that genes belonging to the "cellular processes and signaling" category are overrepresented in the shifted subset in relation to the remainder of the genome (*P* value < 0.001), with COG groups D (cell cycle control, cell division, chromosome partitioning), U (intracellular trafficking, secretion, and vesicular transport), Z (cytoskeleton), N (cell motility), V (defense mechanisms), and O (post-translational modification, protein turnover, chaperones) more strongly associated with the shifted genes within this category (Fig. S2). As literature is heavily biased toward well-annotated genes, and given that annotation quality often relies on a small subset of experiments which cannot capture the full extent of protein functions (39, 40), we believe that these results might inform neglected aspects of the *P. putida* KT2440 genome that could benefit from efforts to provide further context of their roles in this microbial host.

The interpretation of expression data can be affected by the choice of using either annotation source

Considering that the length of the mRNA transcripts encoded by genes annotated differently between GenBank and RefSeq can vary, we sought to quantify the difference in choosing either reference annotation in the quantity and quality of differentially expressed genes found by a standard RNAseq pipeline (Materials and Methods). Under this analysis, a scatter plot of $\log_2$ fold change values for the majority of the genes that are annotated in the same genomic coordinates in the GenBank and RefSeq forms a straight line along the center of the plot, suggesting that the calculated expression levels for genes of the same predicted length on both annotations tend to stay the same. Alternatively, genes with differently sized transcript sequences appear as a noisy cluster scattered across both axes, as illustrated in the four panels of Fig. 4.

Despite this variation, the differences in $\log_2$ fold change between both data sets were not too large, even for shifted transcripts. In fact, the largest variations in log fold change typically belong to transposases (such as PP_3113, PP_3499, PP_3979, PP_4024, and PP_0638), regardless of whether transcript length changes between annotations. Proper transposon identification is a particularly challenging aspect of annotating both prokaryotic and eukaryotic genomes due to their repetitive nature and variable

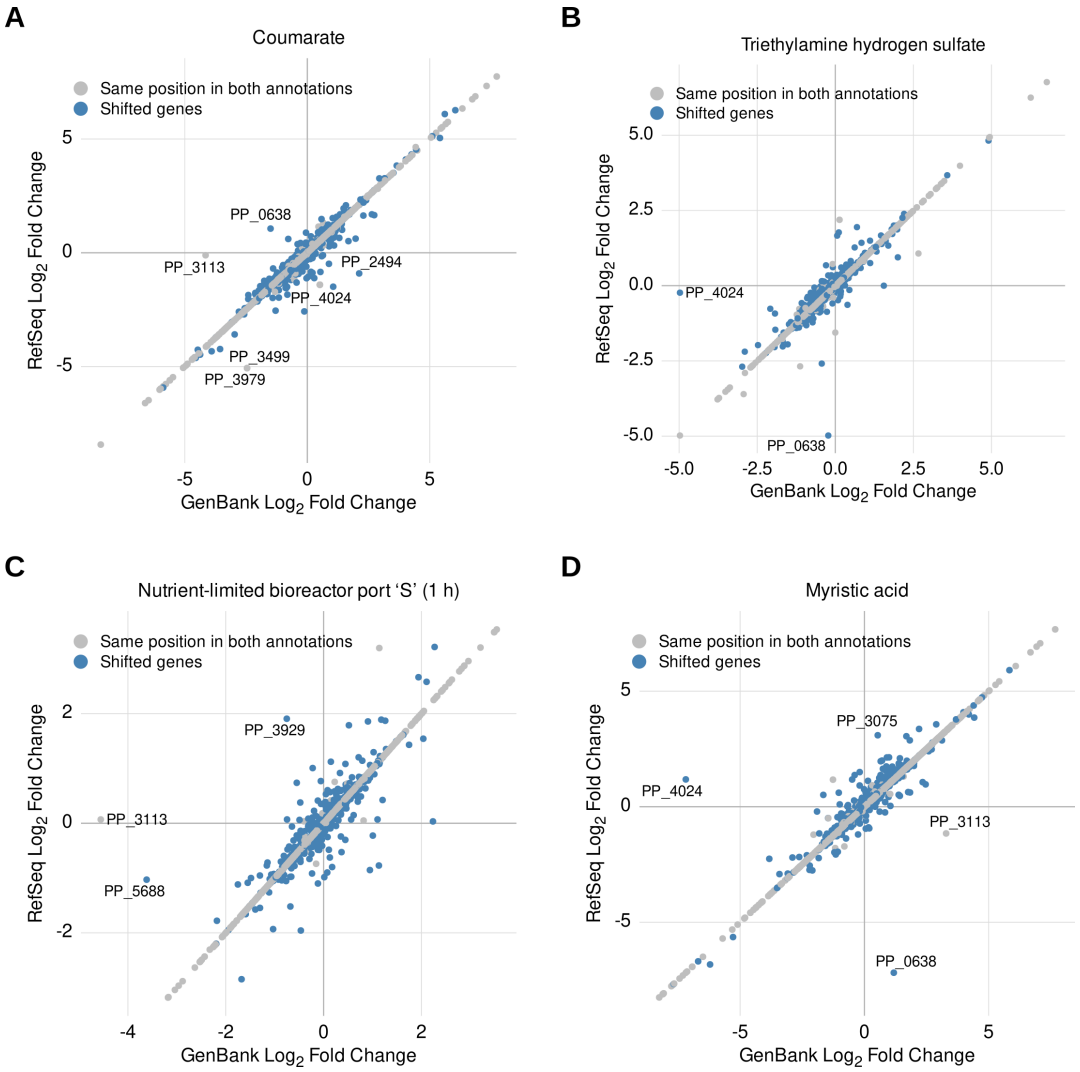

FIG 4 Comparison of calculated $\log_2$ fold change values for shifted and unshifted *P. putida* KT2440 genes using RefSeq or GenBank annotations as a reference to calculate expression data. Panels A–D represent different experiment conditions retrieved from available RNAseq data sets (Table S1). In each panel, blue circles represent shifted genes, whereas gray circles represent unshifted genes. Gene annotations using the conventional "PP_" prefix are given for all genes with an absolute difference greater than 2.5 between their RefSeq and GenBank $\log_2$ fold change values. Among the highlighted genes, only PP_2494 (hypothetical protein), PP_3929 (hypothetical protein), PP_5688 ([2Fe-2S]-binding protein), and PP_3075 (sigma 54-interacting transcriptional regulator), all shifted genes, are not annotated as transposases.

abundance (41–43); therefore, it is unsurprising that this is one of the most prominent changes we found across sources that used different tools to generate gene annotations. Yet, due to the differences between both sources, the subset of genes considered to be significantly differentially expressed can differ between analyses and includes diverse classes of genes (Fig. S3).

While it is beyond the scope of this work to benchmark different RNAseq analysis approaches, at a minimum, we can conclude from these results that a gene can be classified as differentially expressed or not when either reference set of gene annotations is used to calculate the abundances of transcripts from RNAseq data. We did not find examples of genes with altered significance status that were explicitly linked to validated phenotypes or discussed in the original publications we used as a source for this study. However, as publications rely more on the automatic processing of large amounts of data, these results reinforce that robust statistical analysis coupled with the use of sensible thresholds, ideally aided by experimental validation, remains a critical step for

drawing meaningful biological conclusions from large data sets while avoiding spurious correlations inherent to them.

DISCUSSION

In 2019, Ghatak et al. introduced the term "y-ome" as the set of genes of an organism without experimental evidence of function, and described the y-ome of the textbook bacterium *Escherichia coli* K-12 MG1655 after curating data from several online resources (44). However, separating the "known" from the "unknown" relies on foundational assumptions about genes and genomic annotations that are often untested, as researchers are often more interested in biological phenomena specific to their study. Here, we found that not only was the quality of functional annotations uneven for *P. putida* KT2440 in available knowledge bases, but that the existence of two concurring major genomic annotations for this organism creates a divide in the literature (Fig. 1), which, if unaccounted for, contributes to variable interpretation of data and confusion due to imprecision in describing specific gene loci. For example, when validating a set of mutants for fatty acid and alcohol metabolism, Thompson et al. generated a *P. putida* KT2440 strain with a complete internal in-frame deletion of PP_2675, a gene that encodes a type of cytochrome C protein involved in the assimilation of alcohols (45). In the GenBank version of the genome, the start codon of the downstream gene PP_2676 lies within the ORF of PP_2675, meaning that the resulting strain is an unintentional double mutant for both PP_2675 and PP_2676. On the other hand, in the RefSeq annotation, the PP_2676 (PP_RS13935) start codon lies six base pairs downstream to the end of PP_2675 (PP_RS13930), and there is no overlap between the transcripts. We found this scenario to occur at over one hundred loci across our data set (File S3), and while the specific case of PP_2675 was acknowledged in a correction issued shortly after this first publication (46), determining which version of the ORFs corresponds to the biologically expressed *in vivo* remains challenging since they both derive from *in silico* predictions.

Such discrepancies require additional experimental efforts to fully characterize phenotypic behavior in *P. putida* KT2440 under different experimental conditions, which, in turn, hinders progress in elucidating new biological functions in this host. Moreover, since common databases provide different versions of the *P. putida* KT2440 genome annotation, obtaining a coherent picture of the interlinked knowledge of metabolic and physiological traits of this organism can be challenging. As a result, intense data-driven approaches that rely on information available from different sources, like large-scale metabolic or genomic models built by different research groups, can produce conflicting results that require extensive manual curation to be reconciled (6, 47). Despite further published examples addressing these differences being hard to locate, 897 genes, over 16% of all of the annotated genes in *P. putida* KT2440, fall under the shifted category (Fig. 2). As this subgroup is biased toward specific functional annotation categories (Fig. 3), this effect may permeate research on KT2440. This is especially relevant considering that the broader context surrounding start codons can heavily influence protein expression in bacteria (48, 49). If shifted genes are selected for heterologous expression, such as in metabolic engineering or synthetic biology approaches, these variations may lead to differing biological outcomes that are solely dependent on the data set used, even if variables, such as the Shine-Dalgarno sequence, are controlled for.

Propagating and harmonizing annotations across databases is not a genomics- or *P. putida*-specific issue but rather a widely discussed challenge across multiple fields (50, 51). For instance, it is well documented that differences in genome assemblies available in GENCODE and annotations from databases, such as RefSeq, Ensembl, and UCSC, have been shown to affect variant calling and transcript quantification in human data sets, especially due to reads spanning exon-exon junctions (52–55). Analyzing prokaryotic expression data sets can present its own sets of methodological challenges, but inconsistencies between databases are often not fully appreciated by researchers working within the narrow scope of their model organisms, even though the inadvertent

use of different references can also lead to varying interpretations of such data sets (Fig. 4). This effect may be amplified when new taxonomic classifications for well-established strains or gene naming conventions, such as the RefSeq prokaryotic re-annotation project, or the recently proposed taxonomic re-classification of *P. putida* KT2440 as a *Pseudomonas alloputida* (56, 57), break linkages of the known literature. In this sense, we believe that isolated attempts at correcting for these issues, however well-intentioned, may only add to the noise of an already growing set of different genomic resources if not consistently adopted and/or frequently updated to accurately reflect the growing body of work available in the literature. Moreover, while we have illustrated this problem for *P. putida* KT2440, a fairly well-described strain, it undoubtedly exists for all microbes where genome annotations are updated asynchronously from the research community. While a proper solution for this might require community-level effort, establishing and upholding a coherent set of good practices for publications utilizing and reporting information derived from mainstream gene annotation repositories can be a strategy for mitigating some of the current confusion. For instance, informing retrieval dates in addition to the source repository and accession codes of annotations used in a publication is a simple step to bolster reproducibility and account for frequent database updates. All this considered, we hope the present work raises awareness of a latent but underappreciated problem in the *P. putida* KT2440 genome annotation landscape, which should encourage researchers to deliberately choose, and faithfully describe, their specific reference materials in publications.

ACKNOWLEDGMENTS

We thank our colleagues in the MetaGenLab and m-group research groups, as well as Tiago Cabral Borelli, M.Sc., for their constructive comments on this work.

This work was supported by the São Paulo Research Foundation (FAPESP, awards # 2021/01748-5 and 2019/25432-7). M.-E.G. was supported by CNPq Research Productivity Scholarship (award # 304089/2023-0). T.E. and A.M. acknowledge funding and support by the Joint BioEnergy Institute, U.S. Department of Energy, Office of Science, Biological and Environmental Research Program under Award Number DE-AC02-05CH11231 with Lawrence Berkeley National Laboratory. The work conducted at the Joint BioEnergy Institute was supported by the U.S. Department of Energy, Office of Science, Biological and Environmental Research Program, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy.

Conceptualization: G.M.V.D.S. Data acquisition, processing, and interpretation: G.M.V.S. Preparation of figures: G.M.V.D.S. Initial draft of manuscript: G.M.V.D.S. and T.E. Editing of manuscript: T.E., G.M.V.D.S., A.M., and M.-E.G. Raised funds: M.-E.G. and A.M. All authors have read, given feedback, and approved the manuscript for publication.

AUTHOR AFFILIATIONS

[1]Department of Biology, Faculty of Philosophy, Sciences and Letters at Ribeirão Preto, University of São Paulo, Ribeirão Preto, State of São Paulo, Brazil
[2]The Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Emeryville, California, USA
[3]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[4]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA

AUTHOR ORCIDs

Guilherme Marcelino Viana de Siqueira http://orcid.org/0000-0002-3645-6363
Thomas Eng http://orcid.org/0000-0002-4974-3863
Aindrila Mukhopadhyay http://orcid.org/0000-0002-6513-7425

María-Eugenia Guazzaroni http://orcid.org/0000-0002-4657-3731

FUNDING

Funder	Grant(s)	Author(s)
Fundação de Amparo à Pesquisa do Estado de São Paulo	2021/01748-5	Guilherme Marcelino Viana de Siqueira María-Eugenia Guazzaroni
Fundação de Amparo à Pesquisa do Estado de São Paulo	2019/25432-7	Guilherme Marcelino Viana de Siqueira
Conselho Nacional de Desenvolvimento Científico e Tecnológico	304089/2023-0	María-Eugenia Guazzaroni
U.S. Department of Energy	DE-AC02-05CH11231	Thomas Eng Aindrila Mukhopadhyay

AUTHOR CONTRIBUTIONS

Guilherme Marcelino Viana de Siqueira, Conceptualization, Formal analysis, Investigation, Validation, Visualization, Writing – original draft, Writing – review and editing | Thomas Eng, Writing – original draft, Writing – review and editing | Aindrila Mukhopadhyay, Funding acquisition, Writing – review and editing | María-Eugenia Guazzaroni, Funding acquisition, Writing – review and editing

DATA AVAILABILITY

The source code used for analysis and figure generation in this work is available in GitHub repositories (see Materials and Methods). The BioProject accession codes for all RNA-seq data sets analyzed in this work are listed in Table S1.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

File S1 (mSphere00391-25-s0001.txt). Metadata for *P. putida* publication search results.
File S2 (mSphere00391-25-s0002.txt). List of shifted genes.
File S3 (mSphere00391-25-s0003.txt). List of genes with different overlap status between annotations.
Supplemental figures and tables (mSphere00391-25-s0004.pdf). Figures S1 to S3; Tables S1 and S2.
Supplemental file legends (mSphere00391-25-s0005.pdf). Legends for Files S1 to S3.

REFERENCES

1. de Lorenzo V, Pérez-Pantoja D, Nikel PI. 2024. *Pseudomonas putida* KT2440: the long journey of a soil-dweller to become a synthetic biology chassis. J Bacteriol 206:e00136-24. https://doi.org/10.1128/jb.00136-24
2. Kampers LFC, Volkers RJM, Martins Dos Santos VAP. 2019. *Pseudomonas putida* KT2440 is HV1 certified, not GRAS. Microb Biotechnol 12:845–848. https://doi.org/10.1111/1751-7915.13443
3. Nikel PI, de Lorenzo V. 2018. *Pseudomonas putida* as a functional chassis for industrial biocatalysis: from native biochemistry to trans-metabolism. Metab Eng 50:142–155. https://doi.org/10.1016/j.ymben.2018.05.005
4. Nelson KE, Weinel C, Paulsen IT, Dodson RJ, Hilbert H, Martins dos Santos VAP, Fouts DE, Gill SR, Pop M, Holmes M, et al. 2002. Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. Environ Microbiol 4:799–808. https://doi.org/10.1046/j.1462-2920.2002.00366.x
5. Nogales J, Mueller J, Gudmundsson S, Canalejo FJ, Duque E, Monk J, Feist AM, Ramos JL, Niu W, Palsson BO. 2020. High-quality genome-scale metabolic modelling of *Pseudomonas putida* highlights its broad metabolic capabilities. Environ Microbiol 22:255–269. https://doi.org/10.1111/1462-2920.14843
6. Yuan Q, Huang T, Li P, Hao T, Li F, Ma H, Wang Z, Zhao X, Chen T, Goryanin I. 2017. Pathway-consensus approach to metabolic network reconstruction for *Pseudomonas putida* KT2440 by systematic comparison of published models. PLoS ONE 12:e0169437. https://doi.org/10.1371/journal.pone.0169437
7. Tokic M, Hatzimanikatis V, Miskovic L. 2020. Large-scale kinetic metabolic models of *Pseudomonas putida* KT2440 for consistent design of metabolic engineering strategies. Biotechnol Biofuels 13:33. https://doi.org/10.1186/s13068-020-1665-7
8. Puchałka J, Oberhardt MA, Godinho M, Bielecka A, Regenhardt D, Timmis KN, Papin JA, Martins dos Santos VAP. 2008. Genome-scale reconstruction and analysis of the *Pseudomonas putida* KT2440 metabolic network

facilitates applications in biotechnology. PLoS Comput Biol 4:e1000210. https://doi.org/10.1371/journal.pcbi.1000210
9. Nogales J, Palsson BØ, Thiele I. 2008. A genome-scale metabolic reconstruction of *Pseudomonas putida* KT2440: iJN746 as a cell factory. BMC Syst Biol 2:79. https://doi.org/10.1186/1752-0509-2-79
10. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745. https://doi.org/10.1093/nar/gkv1189
11. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F, Derbyshire MK, Durkin AS, Gonzales NR, Gwadz M, et al. 2021. RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. Nucleic Acids Res 49:D1020–D1028. https://doi.org/10.1093/nar/gkaa1105
12. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 44:6614–6624. https://doi.org/10.1093/nar/gkw569
13. Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, Li W, Chitsaz F, Derbyshire MK, Gonzales NR, Gwadz M, Lu F, et al. 2018. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res 46:D851–D860. https://doi.org/10.1093/nar/gkx1068
14. Belda E, Heck RGA, José Lopez-Sanchez M, Cruveiller S, Barbe V, Fraser C, et al. 2016. The revisited genome of *Pseudomonas putida* KT2440 enlightens its value as a robust metabolic chassis: re-annotation of the *Pseudomonas putida* KT2440 genome. Environ Microbiol 18:3403–3424. https://doi.org/10.1111/1462-2920.13230
15. Winsor GL, Griffiths EJ, Lo R, Dhillon BK, Shay JA, Brinkman FSL. 2016. Enhanced annotations and features for comparing thousands of *Pseudomonas* genomes in the *Pseudomonas* genome database. Nucleic Acids Res 44:D646–53. https://doi.org/10.1093/nar/gkv1227
16. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, Keseler IM, Krummenacker M, Midford PE, Ong Q, Ong WK, Paley SM, Subhraveti P. 2019. The BioCyc collection of microbial genomes and metabolic pathways. Brief Bioinform 20:1085–1093. https://doi.org/10.1093/bib/bbx085
17. Bateman A, Martin M-J, Orchard S, Magrane M, Adesina A, Ahmad S, Bowler-Barnett EH, Bye-A-Jee H, Carpentier D, Denny P, et al. 2025. UniProt: the universal protein knowledgebase in 2025. Nucleic Acids Res 53:D609–D617. https://doi.org/10.1093/nar/gkae1010
18. Kanehisa M. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30. https://doi.org/10.1093/nar/28.1.27
19. Lim HG, Rychel K, Sastry AV, Bentley GJ, Mueller J, Schindel HS, Larsen PE, Laible PD, Guss AM, Niu W, Johnson CW, Beckham GT, Feist AM, Palsson BO. 2022. Machine-learning from *Pseudomonas putida* KT2440 transcriptomes reveals its transcriptional regulatory network. Metab Eng 72:297–310. https://doi.org/10.1016/j.ymben.2022.04.004
20. Rychel K, Decker K, Sastry AV, Phaneuf PV, Poudel S, Palsson BO. 2021. iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. Nucleic Acids Res 49:D112–D120. https://doi.org/10.1093/nar/gkaa810
21. Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, Kuehl JV, Melnyk RA, Lamson JS, Suh Y, Carlson HK, Esquivel Z, Sadeeshkumar H, et al. 2018. Mutant phenotypes for thousands of bacterial genes of unknown function. Nature 557:503–509. https://doi.org/10.1038/s41586-018-0124-0
22. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. Welcome to the tidyverse. JOSS 4:1686. https://doi.org/10.21105/joss.01686
23. Tenenbaum D. 2024. KEGGREST: client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG). Bioconductor. Available from: https://doi.org/10.18129/B9.BIOC.KEGGREST. Retrieved 22 Feb 2025.
24. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. 2021. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res 49:D274–D281. https://doi.org/10.1093/nar/gkaa1018
25. Tennekes M. 2023. treemap: Treemap visualization. Available from: https://doi.org/10.32614/CRAN.package.treemap. Retrieved 27 May 2025.
26. Harrison J. 2022. RSelenium: R bindings for 'Selenium WebDriver'. Available from: https://doi.org/10.32614/CRAN.package.RSelenium. Retrieved 27 May 2025.
27. Avendaño R, Muñoz-Montero S, Rojas-Gätjens D, Fuentes-Schweizer P, Vieto S, Montenegro R, Salvador M, Frew R, Kim J, Chavarría M, Jiménez JI. 2023. Production of selenium nanoparticles occurs through an interconnected pathway of sulphur metabolism and oxidative stress response in *Pseudomonas putida* KT2440. Microb Biotechnol 16:931–946. https://doi.org/10.1111/1751-7915.14215
28. Lim HG, Fong B, Alarcon G, Magurudeniya HD, Eng T, Szubin R, et al. 2020. Generation of ionic liquid tolerant *Pseudomonas putida* KT2440 strains via adaptive laboratory evolution. Green Chem 22:5677–5690. https://doi.org/10.1039/D0GC01663B
29. Pobre V, Graça-Lopes G, Saramago M, Ankenbauer A, Takors R, Arraiano CM, Viegas SC. 2020. Prediction of novel non-coding RNAs relevant for the growth of *Pseudomonas putida* in a bioreactor. Microbiology (Reading) 166:149–156. https://doi.org/10.1099/mic.0.000875
30. D'Arrigo I, Cardoso JGR, Rennig M, Sonnenschein N, Herrgård MJ, Long KS. 2019. Analysis of *Pseudomonas putida* growth on non-trivial carbon sources using transcriptomics and genome-scale modelling. Environ Microbiol Rep 11:87–97. https://doi.org/10.1111/1758-2229.12704
31. Molina-Santiago C, Udaondo Z, Gómez-Lozano M, Molin S, Ramos JL. 2017. Global transcriptional response of solvent-sensitive and solvent-tolerant *Pseudomonas putida* strains exposed to toluene. Environ Microbiol 19:645–658. https://doi.org/10.1111/1462-2920.13585
32. Bojanovič K, D'Arrigo I, Long KS. 2017. Global transcriptional responses to osmotic, oxidative, and imipenem stress conditions in *Pseudomonas putida*. Appl Environ Microbiol 83:e03236-16. https://doi.org/10.1128/AEM.03236-16
33. Peng J, Miao L, Chen X, Liu P. 2018. Comparative transcriptome analysis of *Pseudomonas putida* KT2440 revealed its response mechanisms to elevated levels of zinc stress. Front Microbiol 9:1669. https://doi.org/10.3389/fmicb.2018.01669
34. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170
35. Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinform 12. https://doi.org/10.1186/1471-2105-12-323
36. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/s13059-014-0550-8
37. National Center for Biotechnology Information. Prokaryotic RefSeq Genome Re-annotation Project. Available from: https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/reannotation/. Retrieved 24 May 2025. Accessed May 24, 2025
38. Phaneuf PV, Gosting D, Palsson BO, Feist AM. 2019. ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. Nucleic Acids Res 47:D1164–D1171. https://doi.org/10.1093/nar/gky983
39. Haynes WA, Tomczak A, Khatri P. 2018. Gene annotation bias impedes biomedical research. Sci Rep 8:1362. https://doi.org/10.1038/s41598-018-19333-x
40. Schnoes AM, Ream DC, Thorman AW, Babbitt PC, Friedberg I. 2013. Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. PLoS Comput Biol 9:e1003063. https://doi.org/10.1371/journal.pcbi.1003063
41. Hoen DR, Hickey G, Bourque G, Casacuberta J, Cordaux R, Feschotte C, Fiston-Lavier A-S, Hua-Van A, Hubley R, Kapusta A, Lerat E, Maumus F, Pollock DD, Quesneville H, Smit A, Wheeler TJ, Bureau TE, Blanchette M. 2015. A call for benchmarking transposable element annotation methods. Mob DNA 6:13. https://doi.org/10.1186/s13100-015-0044-6
42. O'Neill K, Brocks D, Hammell MG. 2020. Mobile genomics: tools and techniques for tackling transposons. Phil Trans R Soc B 375:20190345. https://doi.org/10.1098/rstb.2019.0345
43. Platt RN, Blanco-Berdugo L, Ray DA. 2016. Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome Biol Evol 8:403–410. https://doi.org/10.1093/gbe/evw009
44. Ghatak S, King ZA, Sastry A, Palsson BO. 2019. The y-ome defines the 35% of *Escherichia coli* genes that lack experimental evidence of function. Nucleic Acids Res 47:2446–2454. https://doi.org/10.1093/nar/gkz030
45. Thompson MG, Incha MR, Pearson AN, Schmidt M, Sharpless WA, Eiben CB, Cruz-Morales P, Blake-Hedges JM, Liu Y, Adams CA, Haushalter RW,

Krishna RN, Lichtner P, Blank LM, Mukhopadhyay A, Deutschbauer AM, Shih PM, Keasling JD. 2020. Fatty acid and alcohol metabolism in *Pseudomonas putida*: functional analysis using random barcode transposon sequencing. Appl Environ Microbiol 86:e01665-20. https://doi.org/10.1128/AEM.01665-20
46. Thompson MG, Incha MR, Pearson AN, Schmidt M, Sharpless WA, Eiben CB, Cruz-Morales P, Blake-Hedges JM, Liu Y, Adams CA, Haushalter RW, Krishna RN, Lichtner P, Blank LM, Mukhopadhyay A, Deutschbauer AM, Shih PM, Keasling JD. 2021. Correction for Thompson et al., "Fatty acid and alcohol metabolism in *Pseudomonas putida*: functional analysis using random barcode transposon sequencing". Appl Environ Microbiol 87:e00177-21. https://doi.org/10.1128/AEM.00177-21
47. Brixi G, Durrant MG, Ku J, Poli M, Brockman G, Chang D, Gonzalez GA, King SH, Li DB, Merchant AT, et al. 2025. Genome modeling and design across all domains of life with Evo 2. bioRxiv. https://doi.org/10.1101/2025.02.18.638918
48. Lipońska A, Ousalem F, Aalberts DP, Hunt JF, Boël G. 2019. The new strategies to overcome challenges in protein production in bacteria. Microb Biotechnol 12:44–47. https://doi.org/10.1111/1751-7915.13338
49. Saito K, Green R, Buskirk AR. 2020. Translational initiation in *E. coli* occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing. elife 9:e55002. https://doi.org/10.7554/eLife.55002
50. de Crécy-Lagard V, Amorin de Hegedus R, Arighi C, Babor J, Bateman A, Blaby I, Blaby-Haas C, Bridge AJ, Burley SK, Cleveland S, et al. 2022. A roadmap for the functional annotation of protein families: a community perspective. Database (Oxford) 2022:baac062. https://doi.org/10.1093/database/baac062
51. Danchin A, Ouzounis C, Tokuyasu T, Zucker J-D. 2018. No wisdom in the crowd: genome annotation in the era of big data - current status and future prospects. Microb Biotechnol 11:588–605. https://doi.org/10.1111/1751-7915.13284
52. Ungar RA, Goddard PC, Jensen TD, Degalez F, Smith KS, Jin CA, Undiagnosed Diseases Network, Bonner DE, Bernstein JA, Wheeler MT, Montgomery SB. 2024. Impact of genome build on RNA-seq interpretation and diagnostics. Am J Hum Genet 111:1282–1300. https://doi.org/10.1016/j.ajhg.2024.05.005
53. Li H, Dawood M, Khayat MM, Farek JR, Jhangiani SN, Khan ZM, Mitani T, Coban-Akdemir Z, Lupski JR, Venner E, Posey JE, Sabo A, Gibbs RA. 2021. Exome variant discrepancies due to reference-genome differences. Am J Hum Genet 108:1239–1250. https://doi.org/10.1016/j.ajhg.2021.05.011
54. Zhang Q, Shao M. 2023. Transcript assembly and annotations: bias and adjustment. PLoS Comput Biol 19:e1011734. https://doi.org/10.1371/journal.pcbi.1011734
55. Zhao S, Zhang B. 2015. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics 16:97. https://doi.org/10.1186/s12864-015-1308-8
56. Keshavarz-Tohid V, Vacheron J, Dubost A, Prigent-Combaret C, Taheri P, Tarighi S, Taghavi SM, Moënne-Loccoz Y, Muller D. 2019. Genomic, phylogenetic and catabolic re-assessment of the *Pseudomonas putida* clade supports the delineation of *Pseudomonas alloputida* sp. nov., *Pseudomonas inefficax* sp. nov., *Pseudomonas persica* sp. nov., and *Pseudomonas shirazica* sp. nov. Syst Appl Microbiol 42:468–480. https://doi.org/10.1016/j.syapm.2019.04.004
57. Morimoto Y, Tohya M, Aibibula Z, Baba T, Daida H, Kirikae T. 2020. Re-identification of strains deposited as *Pseudomonas aeruginosa*, *Pseudomonas fluorescens* and *Pseudomonas putida* in genbank based on whole genome sequences. Int J Syst Evol Microbiol 70:5958–5963. https://doi.org/10.1099/ijsem.0.004468