

RESEARCH ARTICLE

Open Access



Combining multiple functional annotation tools increases coverage of metabolic annotation

Marc Griesemer^{1†}, Jeffrey A. Kimbrel^{1†}, Carol E. Zhou², Ali Navid¹ and Patrik D'haeseleer^{1,2*} 

Abstract

Background: Genome-scale metabolic modeling is a cornerstone of systems biology analysis of microbial organisms and communities, yet these genome-scale modeling efforts are invariably based on incomplete functional annotations. Annotated genomes typically contain 30–50% of genes without functional annotation, severely limiting our knowledge of the “parts lists” that the organisms have at their disposal. These incomplete annotations may be sufficient to derive a model of a core set of well-studied metabolic pathways that support growth in pure culture. However, pathways important for growth on unusual metabolites exchanged in complex microbial communities are often less understood, resulting in missing functional annotations in newly sequenced genomes.

Results: Here, we present results on a comprehensive reannotation of 27 bacterial reference genomes, focusing on enzymes with EC numbers annotated by KEGG, RAST, EFICAZ, and the BRENDA enzyme database, and on membrane transport annotations by TransportDB, KEGG and RAST. Our analysis shows that annotation using multiple tools can result in a drastically larger metabolic network reconstruction, adding on average 40% more EC numbers, 3–8 times more substrate-specific transporters, and 37% more metabolic genes. These results are even more pronounced for bacterial species that are phylogenetically distant from well-studied model organisms such as *E. coli*.

Conclusions: Metabolic annotations are often incomplete and inconsistent. Combining multiple functional annotation tools can greatly improve genome coverage and metabolic network size, especially for non-model organisms and non-core pathways.

Keywords: Genome annotation, Functional annotation, Enzyme prediction, Transport prediction, Metabolic modeling

Background

In the early days of genome sequencing, functional annotation involved computational prediction of gene function coupled with extensive manual curation by teams of experts [1–3]. Today, with the exponential explosion of DNA sequencing [4] the fraction of genes that have undergone any degree of manual curation or even experimental validation is becoming vanishingly small [5, 6]. Automated gene annotation tools employing different methodologies with minimal manual curation are widely used, functionally annotating by homology to

existing annotations, or by identification of conserved domains/motifs within a coding sequence [7, 8]. Many draft genomes and metagenome bins are often run through a single annotation pipeline where genome annotations are inherited from previous genome annotations. Even when multiple annotation tools are used, integrating the different outputs in a cohesive manner remains a major challenge [9, 10]. However, individual annotation tools often return annotations for different subsets of genes, offering the potential to greatly increase the coverage of annotations by combining the outputs of multiple tools if the barrier for integration can be overcome. Accurate and comprehensive genome annotations are essential for deciphering evolutionary, systematic and ecosphere functions of sequenced organisms. In particular, constraint-based metabolic modeling has emerged as a key tool in systems biology [11].

* Correspondence: dhaeseleer2@llnl.gov

[†]Marc Griesemer and Jeffrey Kimbrel contributed equally to this work.

¹Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA

²Global Security Computing Applications Division, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA



Genome-scale metabolic models implicitly assume complete and accurate functional annotation. However, 30–50% of genes in a typical genome still lack any functional annotation [12], a statistic which has not improved much over the past two decades of genome sequencing [3]. More than 30% of these unannotated genes are estimated to have metabolic functions [13] leaving a significant gap in our understanding of the underlying metabolic processes. In addition, annotated genes include a large (and potentially growing) fraction of misannotations [14]. Further, high-throughput untargeted metabolomics often contain a large fraction of peaks that cannot be reliably matched to any known metabolites [15], and many of those that can be identified often do not match any metabolic reconstruction of the microbial species involved [16], providing another strong indication of the extent of microbial metabolism we are missing with current metabolic annotation methods. As the scale and complexity of the biosystems that we study increases, there is a need to examine the metabolic interactions between constituent organisms and assess the system-level outcome of these exchanges. Metabolic modeling efforts, are moving beyond studying core metabolic pathways in a single organism towards multi-species models, real-world communities and ecosystems [17–19], and incorporation of complex ‘omics and metabolite data, emphasizing the need for a more complete coverage of the metabolic functions identified in microbial genomes, a much wider range of metabolites found in complex communities, and a greater emphasis on transport and exchange processes.

In constraint-based genome-scale modeling methods such as Flux Balance Analysis [11], the issue of missing metabolic annotations is dealt with by “gap filling” - the addition of a set of metabolic reactions beyond those that were derived directly from the genome annotation [20]. A variety of gap filling algorithms have been developed to predict the missing reactions necessary to make the metabolic network model sufficiently complete to produce biomass [21–24]. In a broad collection of 130 genome-scale metabolic models added to the ModelSEED database [25], on average 56 additional gap filled reactions were needed for each model to produce biomass in simple defined nutrient media. Even after those additions an average of one-third of the reactions in each model were still blocked, meaning that there were still enough reactions missing in the network to preclude metabolic flux through those reactions [25, 26]. In addition, the number of reactions that can partake in a gap filling solution is vast (3270 in the case of *E. coli*), and the sets of reactions generated by different gap filling algorithms may have little or no overlap with each other [27]. Clearly, a more complete identification and annotation of metabolic reactions would be preferable to

the addition of dozens of poorly supported reactions just to patch the holes in the network.

Recent genome-scale modeling of *Clostridium beijerinckii* NCIMB 8052 [28] demonstrated that the total number of genes and reactions included in the final curated model could be almost doubled by incorporating multiple annotation tools. The reconstruction of the *C. beijerinckii* metabolic network used 3 different database sources (SEED [29], KEGG [30], and RefSeq annotations captured in BioCyc [31]) to evaluate annotation coverage and produce a more robust model. Each annotation source contributed only about half of the reactions in the final curated model. Only a third of the reactions were present in all three sources, and these reactions were found to contribute significantly to a core set of active reactions in validation simulations. The small overlap between annotations was not simply due to any source contributing more heavily to a particular area of metabolism, nor did any one source outperform another in terms of model connectivity. Likewise, in an analysis of nine prokaryotic genomes, the three enzyme annotation sources used – NCBI, KEGG, and the PEDANT protein database [32] – only agreed on less than one third of the annotated genes [33].

Accurate metabolic models also rely on accurate determination of substrate transport between the bacterium and its environment. Transporter annotations have rarely been used in genome scale metabolic modeling because of the difficulty in computationally determining the exact substrate being transported. Because of this, many metabolic modeling methods simply assume that a transporter exists for the import of any necessary metabolite - an assumption that is incorrect in some cases. For example, the yogurt bacterium *Streptococcus thermophilus* has an unusual growth phenotype in that it grows poorly on glucose, even though it possesses all the required metabolic enzymes [34]. Instead it preferably imports the disaccharide lactose, hydrolyzes it to glucose and galactose, then secretes the galactose back out of the cell. *S. thermophilus* lacks the typical glucose phosphotransferase system used by many bacteria, and instead has an efficient lactose import mechanism that makes it well adapted to grow in milk [34]. Prediction tools such as TransportDB's Transporter Automatic Annotation Pipeline (TransAAP, [35]) now allow researchers to generate substrate predictions that are sufficiently detailed to be included in metabolic pathways, and could give insights into growth or metabolite exchange phenotypes that are not readily apparent from the metabolic pathways present in the genome.

We undertook an investigation into the effectiveness of several popular tools for genome annotation and their overlap with each other, focusing specifically on enzymatic annotations characterized by EC (Enzyme Commission)

numbers, because those can be most unambiguously mapped across annotations from different sources [36]. Due to the current limitations and difficulties of merging the functional ontologies from different annotation tools, for this analysis we focused on those that produce EC annotations. Limiting our analysis to only EC numbers allowed the mapping of functions between RAST, KEGG, EFICAz and BRENDA. Note that many databases such as RAST, KEGG and MetaCyc also curate their own set of metabolic reactions and reaction variants beyond the canonical EC number hierarchy, however, differing reaction identifiers can be much more difficult to compare across the different tools. Using 27 bacterial reference genomes from BioCyc [31] (see Table 1), we evaluated how many genes, EC numbers, and gene-EC annotations were unique or shared with other tools. We also undertook a study of how transporter annotations were handled between RAST [29], KEGG [30], and TransportDB [35],

focusing especially on transporters with detailed substrate annotations. We hypothesized that by combining annotation tools we could alleviate some of the known problems with lack of coverage of metabolic annotations, especially for less well studied organisms, pathways, and transporters.

Results and discussion

Discrepancies in metabolic annotations between different tools

In total, the RAST, KAAS, EFICAz and BRENDA tools produced 47,447 Gene-EC annotations (“gene X codes for an enzyme with EC number Y”) across the 27 reference genomes, for an average of 1757 annotations per genome. The metabolic gene-EC annotations produced by these automated genome-wide annotation tools differed drastically (Fig. 1). Each tool produced on average between 23% (EFICAz) and 48% (BRENDA) unique gene-EC annotations that were not predicted by any of

Table 1 Reference genomes used in this study

Genome Name	Biocyc ID	Phylum	NCBI Accessions	Proteins
<i>Mycobacterium tuberculosis</i> CDC1551	MTBCDC1551	Actinobacteria	AE000516	4189
<i>Mycobacterium tuberculosis</i> H37Rv	MTBH37RV	Actinobacteria	AL123456	4018
<i>Streptomyces coelicolor</i> A3(2)	SCO	Actinobacteria	NC_003888, NC_003903, NC_003904	8152
<i>Bacteroides thetaiotaomicron</i> VPI-5482	BTHE	Bacteroidetes	AE015928, AY171301	4825
<i>Candidatus Cardinium hertigii</i> ^a	CBTQ1	Bacteroidetes	HG422566, CBQZ010000001- CBQZ010000011	739
<i>Synechococcus elongatus</i> PCC 7942	SYNEL	Cyanobacteria	CP000100, CP000101	2661
<i>Listeria monocytogenes</i> 10403S	10403S_RAST	Firmicutes	CP002002	2814
<i>Bacillus anthracis</i> Ames	ANTHRA	Firmicutes	NC_003997, AE017335, AE017336	5602
<i>Bacillus subtilis</i> 168	BSUB	Firmicutes	AL009126	4185
<i>Clostridium saccharoperbutylacetonicum</i> ATCC 27021	CLOSSAC	Firmicutes	CP004121, CP004122	5821
<i>Eubacterium rectale</i> ATCC 33656	EREC	Firmicutes	CP001107	3626
<i>Clostridioides difficile</i> 630	PDIF272563	Firmicutes	AM180355, AM180356	3809
<i>Agrobacterium fabrum</i> C58	AGRO	Proteobacteria	AE008687, AE008688, AE008689, AE008690	5402
<i>Aurantimonas manganoxydans</i> SI85-9A1	AURANTIMONAS	Proteobacteria	AAPJ01000001- AAPJ01000035	3650
<i>Caulobacter crescentus</i> CB15	CAULO	Proteobacteria	AE005673	3737
<i>Caulobacter crescentus</i> NA1000	CAULONA1000	Proteobacteria	CP001340	3885
<i>Escherichia coli</i> CFT073	ECOL199310	Proteobacteria	AE014075	5379
<i>Escherichia coli</i> K-12 substr. W3110	ECOL316407	Proteobacteria	NC_007779	4410
<i>Escherichia coli</i> B str. REL606	ECOL413997	Proteobacteria	CP000819	4209
<i>Escherichia coli</i> K-12 substr. MG1655 ^a	ECOLI	Proteobacteria	U00096	4140
<i>Escherichia coli</i> O157:H7 str. EDL933	ECO0157	Proteobacteria	AE005174, AF074613	5449
<i>Candidatus Evansia muelleri</i> ^b	EVA	Proteobacteria	LM655252	330
<i>Helicobacter pylori</i> 26,695	HPY	Proteobacteria	CP003904	1594
<i>Methylosinus trichosporium</i> OB3b	MOB3B	Proteobacteria	NZ_ADVE02000001- NZ_ADVE02000003	4344
<i>Candidatus Portiera aleyrodidarum</i> BT-QVLC ^b	PABTQVLC	Proteobacteria	CP003867	280
<i>Shigella flexneri</i> 2a str. 2457 T	SHIGELLA	Proteobacteria	AE014073	4068
<i>Vibrio cholerae</i> O1 biovar El Tor str. N16961	VCHO	Proteobacteria	AE003852, AE003853	3828

^aTier 1 Pathway Genome Database (EcoCyc)

^bEndosymbiont with reduced genome

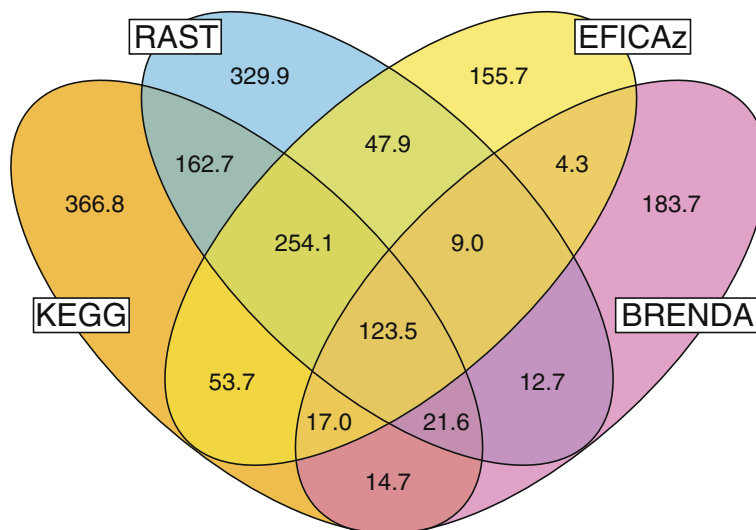


Fig. 1 Large differences exist between the sets of Gene-EC annotations generated by the four annotation tools across the 27 reference genomes

the other tools. Overall, fewer than a quarter of all gene-EC annotations were agreed on by at least 3 tools.

When two annotation tools both assigned a particular gene an EC annotation, the two tools assigned at least one identical EC annotation in more than 50% of cases (Table 2). BRENDA on average had the lowest agreement with other tools (56.0–69.7%). Note also that BRENDA had a larger fraction (47.5%) of gene-EC annotations not shared by any other tools (Fig. 1). In contrast, EFICAz showed the highest agreement with other tools (69.7–86.4%) and had the lowest number of gene-EC annotations not shared by other tools (23.4%).

The denominator is the number of genes across the 27 reference genomes that are covered by both tools. The numerator counts the number of such genes for which both tools provide at least one identical gene-EC number annotation.

Comparing annotation tools against each other can give a sense of which tools are closest to a consensus annotation, or which tools seem to be outliers, however assessing the integrity of these predictions is difficult without experimental validation. Therefore, to determine which tool provides the best ratio of true/false annotation predictions

we compared their predictions to the EcoCyc database [37]. EcoCyc is a gold-standard continuously updated database of experimentally determined and extensively hand-curated enzymatic functions in *Escherichia coli* K-12 substr. MG1655, the most-studied model organism in modern biology. We used the gene-EC numbers annotated in EcoCyc as a set of true positives to evaluate how well the two most commonly used automated annotation tools, RAST and KEGG, are able to assign function to the enzymes in *E. coli* K-12. Overall, there was a high degree of overlap between the RAST and KEGG predictions (Fig. 2) with EcoCyc, however neither tool covered all of EcoCyc, and both tools predicted a small number of reactions not experimentally validated.

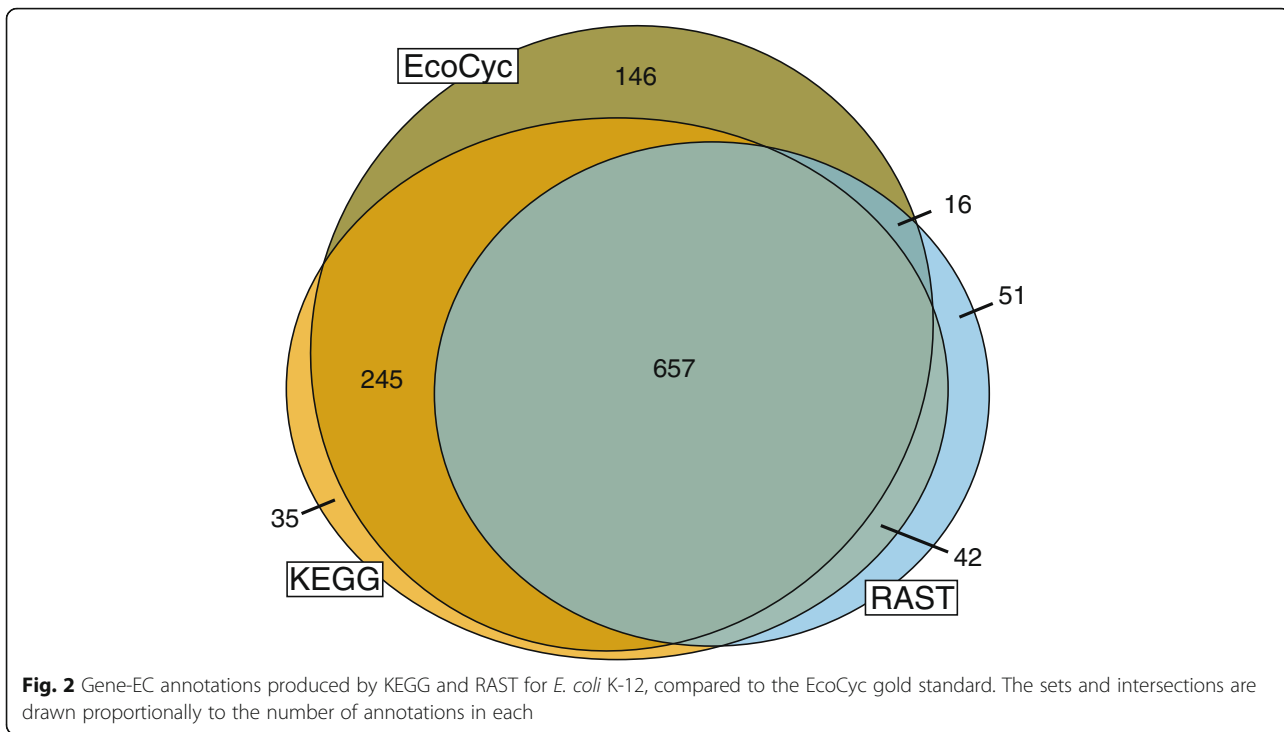
One major caveat of using *E. coli* to evaluate the quality of annotation tools is that so much of our knowledge of microbial metabolism is based on *E. coli*, and therefore annotation tools can be expected to be trained or optimized on *E. coli* to some extent, so performance on *E. coli* is not necessarily indicative of results on other organisms. For example, the KEGG annotation provided by KAAS is done by calculating bidirectional best BLAST hits against annotated reference genomes including *E. coli*, essentially providing a direct lookup of *E. coli* annotations in the KEGG database.

Table 2 Percentage of gene-EC annotation agreements that exist between pairs of tools

Tool Combination	Gene-EC Agreements
KEGG-RAST	16,697/20,915 (79.8%)
KEGG-EFICAz	14,413/16,677 (86.4%)
KEGG-BRENDA	3777/6748 (56.0%)
RAST-EFICAz	12,977/15,694 (82.7%)
RAST-BRENDA	3907/6288 (62.1%)
EFICAz-BRENDA	3902/5601 (69.7%)

Coverage of the metabolic network reconstruction

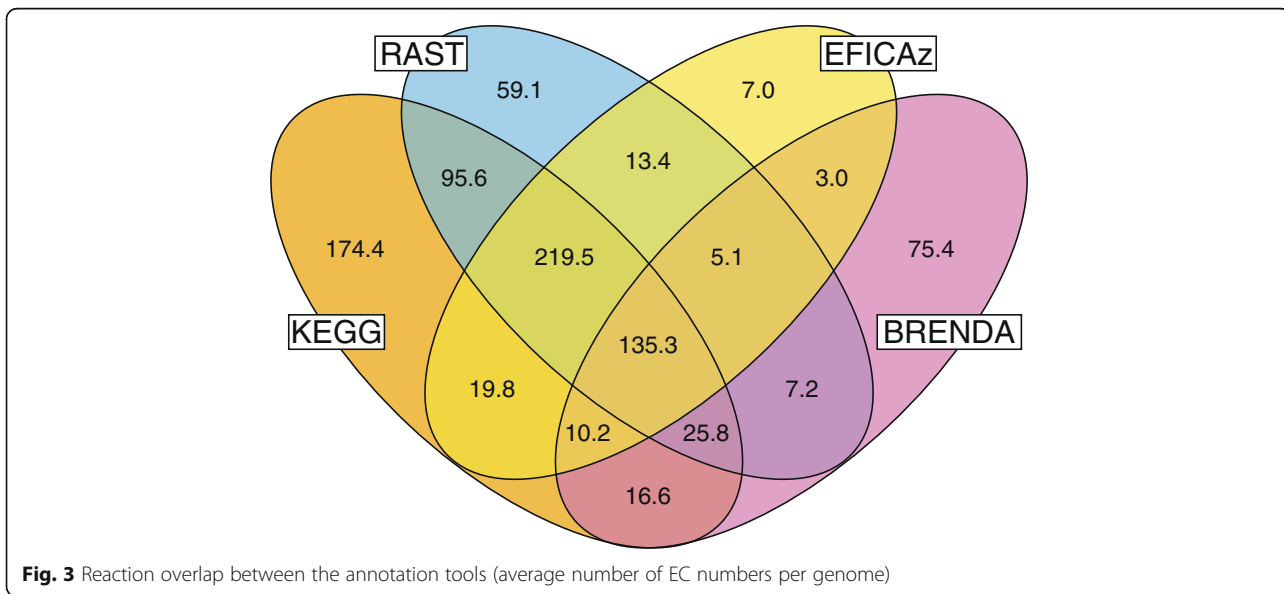
While the individual gene-EC annotations examined in the previous section reflected the quality and agreement between annotation tools, the total set of EC reactions annotated for a genome by each tool reflects the size and coverage of its metabolic network reconstruction. In this case, we simply counted the total number of different EC numbers, regardless of whether multiple genes



are annotated with the same EC number (isozymes), or whether genes were annotated with multiple EC numbers (multifunctional enzymes). On average, the four tools combined produced 868 EC reactions per genome, with the largest agreement between RAST and KEGG (Fig. 3). In general, KEGG produced a larger number of unique EC numbers, which could indicate more over-prediction, or more comprehensive pathway coverage. Note that both RAST and KEGG also generate many reactions without official EC numbers, so in some

cases these annotation tools may produce annotations that are minor variants or subsets of the canonical EC number reaction in EcoCyc.

EFICAz produced the least number of unique EC numbers but had high agreement with RAST and KEGG (86% of EFICAz EC numbers were also generated by both RAST and KEGG – see Fig. 3), even though it uses very different annotation methods, which suggests it can be helpful in validating those annotation tools. Note that EFICAz also produces incomplete “three-digit” EC number



annotations (e.g. 1.2.3.-) which may be useful for hole filling but were not considered in this analysis.

BLASTing against the BRENDA database of reference enzymes produced the smallest number of annotations, but a high fraction of unique EC numbers. Interestingly, of the top 10 unique EC numbers produced by this method, only one is also covered by RAST and KEGG, two of the EC numbers have been deprecated by the Enzyme Commission, and six are EC numbers that have been assigned in 2000 or later and may not have been incorporated into the predictions by the other annotation tools yet. So even though a simple BLAST against a reference database such as BRENDA proves to be one of the less effective means for assigning metabolic functions (compared to the more sophisticated HMM or protein family based methods used by the other tools), it may still have some value to capture recently described enzymes not already covered by the other tools.

While counting the number of EC numbers reflects the size of the metabolic network, counting the numbers of genes that have received any metabolic annotation reflects the genome annotation coverage. Additional file 1: Figure S1 shows the number of genes in all of the reference genomes annotated with one or more EC numbers by each of the tools. On average, 1361 genes per genome were assigned a function with at least one tool, and more than 65% of these genes were assigned a metabolic function by more than one tool. The results show that just as each tool adds a significant number of reactions to the metabolic network model, each tool also

significantly contributes to the number of genes covered with metabolic annotations.

The EC numbers on which the different tools most often agree across the 27 reference genomes tended to belong to well-studied core metabolic pathways. Out of the 79 EC numbers on which all four tools agreed in at least half of the genomes (Additional file 2), more than three quarters (61/79) were involved in biosynthesis or biodegradation of amino acids, nucleotides, carbohydrates, and cofactors; or in the processing of RNA, DNA and proteins. In contrast, almost none of these EC numbers were involved in biosynthesis or degradation of fatty acids, lipids, aromatics compounds, or secondary metabolites.

The differing sets of annotations produced by each tool can enable a user to trade off confidence for coverage, with higher confidence obtained when accepting only annotations that were agreed upon by multiple tools (the intersection), or higher coverage obtained by using the combined set of annotations from multiple tools (the union). To examine the effect of taking the intersection (higher confidence) or union (higher coverage) of the annotation tools, we compared combinations of the four annotation tools against the *E. coli* K-12 “gold standard” metabolic reactions in EcoCyc (Fig. 4). These combinations included the single tool annotations, as well as the union and intersection of all four tools combined in pairs, triplets and quartets. The resulting EC annotations from these combinations were then compared to the 1064 EC numbers from EcoCyc, and the number of true positives, false positives, and

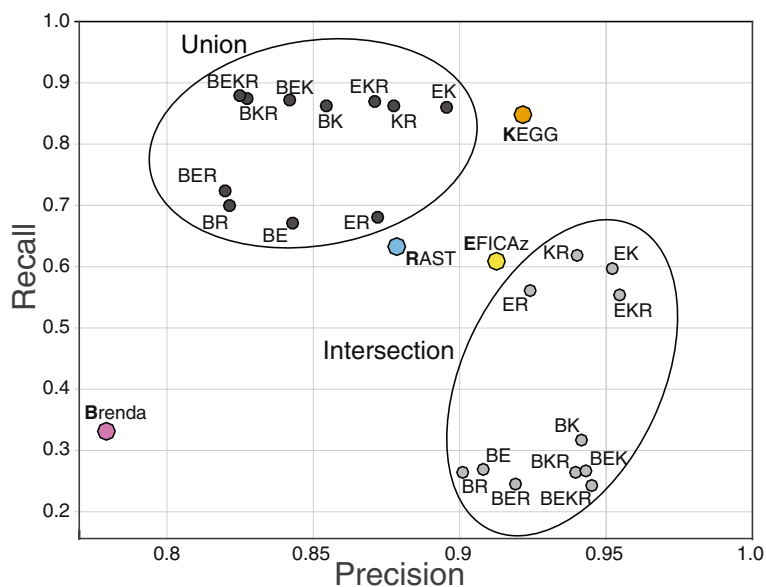


Fig. 4 Precision vs Recall of EC numbers for different combinations of tools on EcoCyc. Individual tools are denoted by B, E, K, or R for BRENDA, EFICAZ, KEGG, and RAST, respectively. For each combination of tools, we calculated precision and recall for both the union and intersection of the sets of EC numbers generated by each tool. The union corresponds to the set of EC numbers generated by at least one of the tools in the combination, while the intersection corresponds to those EC numbers generated by every single tool in the combination

false negatives were calculated for each combination, as well as the resulting Precision and Recall (Table 3).

A more lenient annotation policy (e.g. merging annotations from all tools) will tend to generate fewer false negatives but more false positives, achieving a higher Recall at the expense of lower Precision. Conversely, a more restrictive annotation policy (e.g. only including EC numbers if all tools agree on them) can increase Precision, but at the expense of a lower Recall. Figure 4 shows a plot of precision versus recall for all the different combinations of tools. Out of the individual tools, KEGG performed best in terms of both precision and recall on this dataset (although as mentioned before, performance on *E. coli* K12 may not reflect performance on other genomes), and a simple Blast against the BRENDA database performed worst. Combinations that contain some union of the tools have a higher recall than each of the individual tools in the combination, but a somewhat lower precision (> 80%). In contrast, intersections of annotations from two or more tools show very high (> 90%) precision but much lower recall (< 65%). A consensus annotation that produces both higher recall and higher precision might be achieved by means of a weighted sum of all the annotation sources, similar to the approach taken by EnzymeDetector [33].

Even though we expected the *E. coli* K-12 genome to be a best-case annotation candidate, there were still significant differences in the annotations produced by the different tools, with each tool only covering a subset of the known enzymes in EcoCyc. The four annotation tools annotated a significantly larger fraction of the genome, and showed much more agreement on *E. coli* than on more remote lineages such as Actinomycetes, Bacteroidetes, or Clostridia (Fig. 5a). For *E. coli* K-12, 60% of EC numbers were agreed on by 3 or more tools, while 28% EC numbers come from only a single tool. In contrast, for *C. difficile* 630, only 33% of EC numbers were agreed on by 3 or more tools, and 48% of EC numbers come from only a single tool. Note that this also seems to justify the use of multiple annotation tools in the recent work on modeling of *Clostridium beijerinckii* NCIMB 8052 [28]. Compared to the five *E. coli* strains in our dataset, the annotation tools also cover on average around 30% fewer genes for the 13 genomes in the bottom half of Fig. 5b. We see a similar effect when we

compare the annotation coverage for *B. subtilis* – arguably the best studied Gram-positive model organism – with all 8 other Gram-positive genomes in our dataset. These results suggest that genome coverage for each tool, and agreement in annotations across tools are significantly worse for organisms that are more phylogenetically distant from well-studied model organisms, making it all the more important to combine multiple tools when annotating these genomes.

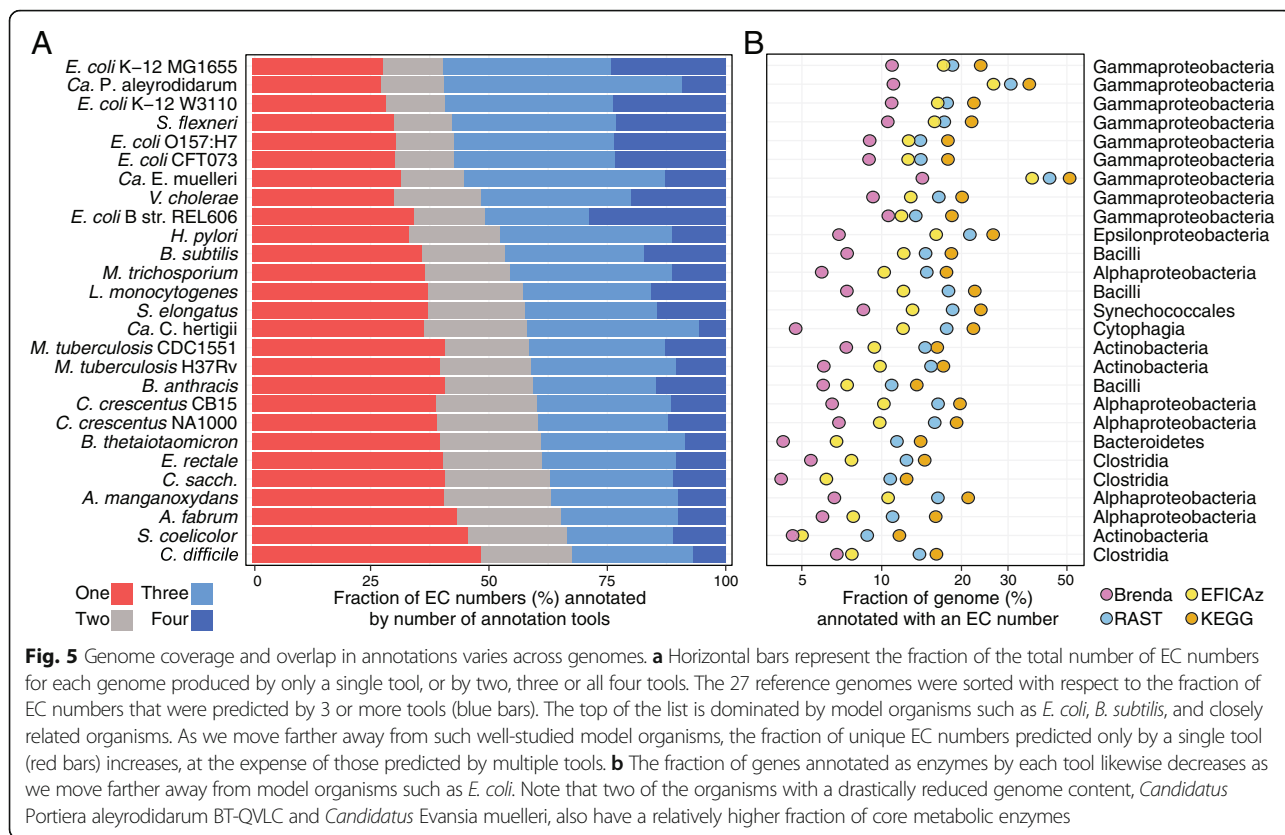
Transporter annotations

Knowledge of the molecules and substrates an organism can transport and exchange with the environment can help to build a more accurate metabolic model. Both RAST and KEGG include membrane transport annotations, yet both tools yielded on average only 114 and 204 transporter predictions per genome, respectively (Fig. 6a and Additional file 1: Figure S2). Many of these annotated transporters lack substrate predictions (52% of transporter annotations in RAST, 25% in KEGG) or have ambiguous substrate predictions (ranks 3–4 (Table 4, Additional file 3); 20% in RAST, 28% in KEGG), while less than half have substrate predictions that are sufficiently detailed to be incorporated in a metabolic model (ranks 1–2; 28% in RAST, 48% in KEGG; Fig. 6b). In contrast, TransportDB produces an average of 426 transport annotations per genome, and most of those have specific substrate predictions (59% rank 1–2; 32% rank 3–4, 10% rank 5; Fig. 6b).

Transporter annotations by RAST, KEGG and TransportDB showed surprisingly little overlap. Out of the more than 15,000 genes annotated as transporters (regardless of substrate prediction), the three tools only agree on 2.8% (423/15,161). Out of those, only 130 genes are annotated by all three tools with a specific substrate prediction (ranks 1–2). When two or more tools provide a sufficiently specific substrate annotation, the substrate annotations tend to agree 85% of the time, even if they may not be identical (for example, one transporter was annotated as “leucine/valine”, “leucine”, and “branched-chain amino acid” by TransportDB, RAST and KEGG respectively). Overall, the detailed transporter annotations by TransportDB’s Transporter Automatic Annotation Pipeline provide a significant advance over more general metabolic annotation tools such as RAST and KEGG.

Table 3 Definitions of Precision, Recall and associated terms

Term	Formula	Definition
True Positive	TP	EC numbers predicted by tools and found in EcoCyc.
False Positive	FP	EC numbers predicted by tools and not found in EcoCyc.
False Negative	FN	EC numbers in EcoCyc but not predicted by tools.
Precision	$TP/(TP + FP)$	Fraction of predicted EC numbers that are in EcoCyc.
Recall	$TP/(TP + FN)$	Fraction of EC numbers in EcoCyc correctly predicted by tools.



Conclusions

This analysis has led us to make recommendations for providing a more comprehensive metabolic genome annotation. We found that a single annotation tool is often insufficient unless one is only interested in core metabolism where different tools often agree. Organisms that are phylogenetically far removed from well-studied model organisms are particularly susceptible, in which case annotation tools will tend to diverge far more. In addition, one can trade off confidence in predictions

versus greater coverage by using the intersection or union of multiple annotation tools. BLASTing against a database of reference sequences is generally an inefficient method for annotating enzymes but may be useful to cover more recently assigned EC numbers not yet included by other tools. Still, all these efforts require manual curation to bring together annotation from multiple sources. More tool development is needed to merge annotations beyond simple EC numbers, and a universal reference database for well-balanced reactions and

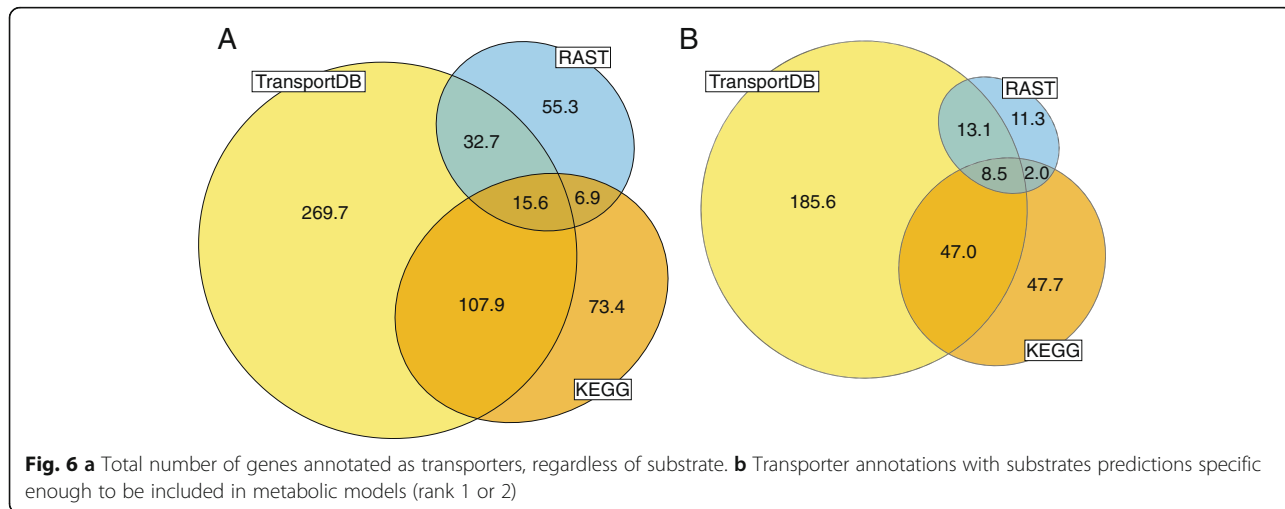


Table 4 Examples of substrate annotation ranking

Rank	Substrate	Examples
1	Metabolite that can be incorporated as a transport reaction in a metabolic model	<ul style="list-style-type: none"> • Fe • lysine
2	Substrate(s) that map to a small number of possible transport reactions	<ul style="list-style-type: none"> • Mg/Co/Ni • aromatic amino acid
3	Broader substrate classes not directly usable to construct a metabolic network	<ul style="list-style-type: none"> • dipeptide • sugar
4	Very broad class of substrates	<ul style="list-style-type: none"> • multidrug efflux • protein
5	No substrate annotated	

Transporter substrates were ranked from most specific (rank 1) to least specific (no substrate, rank 5). See Additional file 3 for the full table

metabolites would be a very valuable resource to merge annotations that use different reaction nomenclatures [38–40]. Likewise, now that annotation tools such as TransportDB are producing significant numbers of transporter annotations with substrate predictions that are precise enough to be included in metabolic modeling, more tool development may be needed to fully take advantage of these substrate predictions in Flux Balance Analysis methods, and move beyond the current implicit assumption used by most algorithms that all metabolites can be transported when needed.

Material and methods

Reference genomes

We focused on a total of 27 genomes of BioCyc Tier 1 & Tier 2 bacteria [31]. These genomes are from a range of phyla, including 15 Proteobacteria, 6 Firmicutes, 3 Actinobacteria, 2 Bacteroidetes and 1 Cyanobacteria. In addition to a range of phyla, these 27 organisms also cover different lifestyles, including a human gut symbiont (*Bacteroides thetaiotaomicron* VPI-5482), pathogens (e.g. *Mycobacterium tuberculosis*), an obligate insect endosymbiont (*Candidatus* *Evansia muelleri*), an unusual manganese oxidizing bacterium (*Aurantimonas manganoxidans* SI85-9A1), and a photoautotrophic cyanobacterium (*Synechococcus elongatus* PCC 7942). Genbank files were downloaded from NCBI (accessions listed in Table 1), and standardized to remove all functional annotation, retaining only the original open reading frames and locus tag/protein identification information (Additional file 4).

Annotation tools

RAST (Rapid Annotation Subsystem Technology, [29]) is an open-source web server for genome annotation, using an assignment propagation strategy based on manually curated subsystems and subsystem-based protein families that automatically guarantees a high degree of assignment consistency. *RAST* returns an analysis of the genes and subsystems in each genome. We used the NMPDR website [41] to generate genome-wide annotations

for our 27 reference genomes and parsed any EC numbers from the functional annotation. This is also the core metabolic annotation tool used by the popular ModelSEED tool for generating draft genome-scale models of metabolism [42].

KEGG (Kyoto Encyclopedia of Genes and Genomes, [30]) is a collection of genome and pathway databases for systems biology. We used KAAS (KEGG Automatic Annotation Server [43, 44]) to generate genome-wide annotations for our 27 reference genomes. KAAS assigns KEGG Orthology (KO) numbers using the bi-directional best hit method (BBH) against a set of default prokaryotic genomes in the KEGG database. We mapped KO numbers to EC numbers using a mapping table provided by the KEGG BRITE Database [45].

EFICAZ [46] uses large scale inference to classify enzymes into functional families, combining 4 methods into a single approach without the need for structural information. Recognition of functionally discriminating residues (FDR) allows *EFICAZ* to use a method called evolutionary footprinting. *EFICAZ* has been rigorously cross validated, and achieves very high precision and recall, even for sequence similarities to known enzymes as low as 40%. We used a local install of *EFICAZ*^{2.5} to generate EC number predictions for our 27 reference genomes.

BRENDA [47] is a database containing 7.2 million enzyme sequences categorized into 82,568 enzymatic functions based on the literature and contains functional and molecular information such as nomenclature, structure, and substrate specificity. We annotated our 27 reference genomes based on a BLAST search at >60% sequence identity against a local copy of the 2011 *BRENDA* database of enzyme reference sequences.

All EC number annotations are available in Additional file 5.

Transport annotations

Where available, pre-generated transporter annotations were downloaded from the TransportDB 2.0 database [35]. For those reference genomes that were not already present in the database (*M. trichosporium* OB3b,

Candidatus C. hertigii, *Candidatus E. muelleri*, and *A. manganoxydans* SI85-9A1), we submitted the predicted protein sequences to the TransAAP web-based transporter annotation tool using the default parameters [48]. We also retrieved transporter annotations from RAST and KEGG. For RAST, annotations were filtered for the subsystem for “Membrane Transport”. For KEGG, we mapped KO numbers to transporter annotations using a mapping table provided by the KEGG BRITE Database [49]. Substrate names were ranked from most to least specific (see Table 3). Substrates that can be incorporated as a transport reaction in a metabolic model were ranked 1 and 2. Broader substrate classes that could be used for gap filling or interpretation of transcriptomics data were ranked 3 and 4, and annotated transporters without substrate prediction were ranked 5. The full table of substrate names and ranking can be found in Additional file 6.

Additional files

Additional file 1: Figure S1. Overlap of annotated genes between the tools (average numbers of genes annotated per genome). **Figure S2.** Average transporter annotations per genome produced by TransportDB (426.0), KEGG (203.8) and RAST (113.7) and the distributions of their substrate specificities (rank 1 is most specific, rank 5 has no substrate prediction). (DOC 3273 kb)

Additional file 2: EC numbers predicted by all 4 tools. Excel file with the EC numbers predicted by all 4 tools in most genomes. (XLS 35 kb)

Additional file 3: Substrate ranks. Excel file with all transporter substrate ranks. (XLS 74 kb)

Additional file 4: Genome data. Zip file with the 27 reference genomes in Genbank format, cleaned to include only coding sequences and locus tags. (ZIP 68984 kb)

Additional file 5: EC annotations. Excel file with all EC annotations by all 4 tools for all 27 genomes. (XLS 3442 kb)

Additional file 6: Transporter annotations. Excel file with all transporter annotations by all 3 tools for all 27 genomes. (XLS 2134 kb)

Acknowledgements

Not applicable

Funding

This work was supported by the Department of Energy through the Genomic Science Program as part of the LLNL Biofuels SFA (contract SCW1039-02). Work at LLNL was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (IM: LLNL-JRNL-750275).

Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files. Accession numbers for the genomes used are included in Table 1.

Authors' contributions

PD conceived of the study. MG, JK, and PD designed the study, processed, and analyzed data. CEZ helped to produce the EFICAZ and BRENDA data. AN provided supervision for the study. MG, JK, and PD wrote the manuscript with input from CEZ and AN. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 20 August 2018 Accepted: 5 November 2018

Published online: 19 December 2018

References

- Kyrpides NC. Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nat Biotechnol.* 2009;27:627–32 Nature Publishing Group.
- Kyrpides NC, Ouzounis CA. Whole-genome sequence annotation: “Going wrong with confidence.”. *Mol Microbiol.* 1999;32:886–7.
- Koonin EV, Mushegian AR, Rudd KE. Sequencing and analysis of bacterial genomes. *Curr Biol.* 1996;6:404–16.
- Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? *PLoS Biol.* 2015;13:e1002195 Public Library of Science.
- Howe D, Costanzo M, Fey P, Gojbori T, Hannick L, Hide W, et al. The future of biocuration. *Nature.* 2008;455:47–50.
- Baumgartner WA Jr, Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L. Manual curation is not sufficient for annotation of genomic databases. 2nd ed. *Bioinformatics.* 2007;23:i41–8.
- Friedberg I. Automated protein function prediction—the genomic challenge. *Brief. Bioinform.* 2006;7:225–42 Oxford University Press.
- Médigue C, Moszer I. Annotation, comparison and databases for hundreds of bacterial genomes. *Res Microbiol.* 2007;158:724–36 Elsevier Masson.
- Ijaq J, Chandrasekharan M, Poddar R, Bethi N, Sundararajan VS. Annotation and curation of uncharacterized proteins- challenges. *Front Genet.* 2015;6:1750.
- Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, et al. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics.* 2015;15:141–61 Springer.
- Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nat Biotechnol.* 2010;28:245–8.
- Hanson AD, Pribat A, Waller JC, de Crécy-Lagard V. ‘Unknown’ proteins and “orphan” enzymes: the missing half of the engineering parts list - and how to find it. *Biochem J.* 2010;425:1–11 Portland Press Limited.
- Ellens KW, Christian N, Singh C, Satagopam VP, May P, Linster CL. Confronting the catalytic dark matter encoded by sequenced genomes. *Nucleic Acids Res.* 2017;45:11495–514 Oxford University Press.
- Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. Valencia A, editor. *PLoS Comp Biol.* 2009;5:e1000605 Public Library of Science.
- da Silva RR, Dorrestein PC, Quinn RA. Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci USA.* 2015;112:12549–50 National Academy of Sciences.
- Bowen BP, Fischer CR, Baran R, Banfield JF, Northen T. Improved genome annotation through untargeted detection of pathway-specific metabolites. *BMC Genomics.* 2011;12:S6 BioMed Central.
- Baran R, Brodie EL, Mayberry-Lewis J, Hummel E, da Rocha UN, Chakraborty R, et al. Exometabolite niche partitioning among sympatric soil bacteria. *Nat Comms.* 2015;6:8289.
- Henry CS, Bernstein HC, Weisenhorn P, Taylor RC, Lee J-Y, Zucker J, et al. Microbial community metabolic modeling: a community data-driven network reconstruction. *J Cell Physiol.* 2016;231:2339–45.
- Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotechnol.* 2017;35:81–9 Nature Publishing Group.
- Thiele I, Palsson BO. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols.* 2010;5:93–121 Nature Publishing Group.

21. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, et al. Systems approach to refining genome annotation. *Proc Natl Acad Sci USA*. 2006;103:17480–4 National Academy of Sciences.
22. Kumar VS, Dasika MS, Maranas CD. Optimization based automated curation of metabolic reconstructions. *BMC Bioinform*. 2007;8:212 BioMed Central.
23. Kumar VS, Maranas CD. GrowMatch: An Automated Method for Reconciling *In Silico/In Vivo* Growth Predictions. Ouzounis CA, editor. *PLoS Comp Biol*. 2009;5:e1000308 Public Library of Science.
24. Benedict MN, Mundy MB, Henry CS, Chia N, Price ND. Likelihood-Based Gene Annotations for Gap Filling and Quality Assessment in Genome-Scale Metabolic Models. Maranas CD, editor. *PLoS Comp Biol*. 2014;10:e1003882 Public Library of Science.
25. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol*. 2010;28:977–82 Nature Publishing Group.
26. Ponce-de-León M, Calle-Espinosa J, Peretó J, Montero F. Consistency Analysis of Genome-Scale Models of Bacterial Metabolism: A Metamodel Approach. Vera J, editor. *PLoS ONE*. 2015;10:e0143626 Public Library of Science.
27. Krumholz EW, IGL L. Sequence-based Network Completion Reveals the Integrality of Missing Reactions in Metabolic Networks. *J Biol Chem*. 2015; 290:19197–207 American Society for Biochemistry and Molecular Biology.
28. Milne CB, Eddy JA, Raju R, Ardekani S, Kim P-J, Senger RS, et al. Metabolic network reconstruction and genome-scale model of butanol-producing strain *Clostridium beijerinckii* NCIMB 8052. *BMC Syst Biol* 2014 8:2. 2011;5:130 BioMed Central.
29. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res*. 2014;42:D206–14 Oxford University Press.
30. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*. 2016;44:D457–62 Oxford University Press.
31. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*. 2016;44:D471–80.
32. Walter MC, Rattei T, Arnold R, Güldener U, Münsterkötter M, Nenova K, et al. PEDANT covers all complete RefSeq genomes. *Nucleic Acids Res*. 2009;37: D408–11 Oxford University Press.
33. Quester S, Schomburg D. EnzymeDetector: an integrated enzyme function prediction tool and database. *BMC Bioinform*. 2011;12:376.
34. Poolman B. Energy transduction in lactic acid bacteria. *FEMS Microbiol Rev*. 1993;12:125–47.
35. Elbourne LDH, Tetu SG, Hassan KA, Paulsen IT. TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic Acids Res*. 2017;45:D320–4.
36. McDonald AG, Tipton KF. Fifty-five years of enzyme classification: advances and difficulties. 2nd ed. *FEBS J*. 2014;281:583–92 Wiley/Blackwell (10.1111).
37. Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, et al. The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res*. 2017;45:D543–50.
38. Bernard T, Bridge A, Morgat A, Moretti S, Xenarios I, Pagni M. Reconciliation of metabolites and biochemical reactions for metabolic networks. *Brief Bioinform*. 2014;15:123–35.
39. Kumar A, Suthers PF, Maranas CD. MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinform*. 2012;13:6 BioMed Central.
40. Lang M, Stelzer M, Schomburg D. BKM-react, an integrated biochemical reaction database. *BMC Biochem*. 2011;12:42 BioMed Central.
41. RAST (Rapid Annotation using Subsystem Technology). The NMPDR, SEED-based, prokaryotic genome annotation service. <http://rast.nmpdr.org>. Accessed 29 Dec 2016.
42. Devoid S, Overbeek R, DeJongh M, Vonstein V, Best AA, Henry C. Automated Genome Annotation and Metabolic Model Reconstruction in the SEED and Model SEED, *Systems Metabolic Engineering*. Totowa: Humana Press; 2013. p. 17–45.
43. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*. 2007;35:W182–5.
44. KEGG Automatic Annotation Server. <http://www.genome.jp/tools/kaas>. Accessed 24 Oct 2016.
45. KEGG BRITE Database. http://www.genome.jp/kegg-bin/get_htext?ko01000.keg. Accessed 29 Nov 2017.
46. Kumar N, Skolnick J. EFICAZ2.5: application of a high-precision enzyme function predictor to 396 proteomes. *Bioinformatics*. 2012;28:2687–8.
47. Placzek S, Schomburg I, Chang A, Jeske L, Ulbrich M, Tillack J, et al. BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res*. 2017;45:D380–8.
48. TransportDB TransAAP. http://www.membranetransport.org/transportDB2/TransAAP_login.html. Accessed 26 Aug 2016.
49. KEGG BRITE Transporters Database. http://www.genome.jp/kegg-bin/get_htext?ko02000.keg. Accessed 16 Aug 2017.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)

