Combining multiple functional annotation tools increases coverage of metabolic annotation

Marc Griesemer[1,§,*], Jeffrey Kimbrel[1,*], Carol Zhou[2], Ali Navid[1], Patrik D'haeseleer[1,2]
[1] Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA
[2] Global Security Computing Applications Division, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA
[§] Corresponding author (griesemer1@llnl.gov)
*Co-first authors

Abstract

Genome-scale metabolic modeling is a cornerstone of computational, omics-enabled systems biology analysis of microbial organisms and communities, yet these genome-scale modeling efforts are invariably based on very incomplete functional annotations. Annotated genomes typically contain 30-50% of genes with little or no 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 (and thus well-annotated) metabolic pathways that are sufficient for growth in pure culture. However secondary metabolic pathways or pathways that are important for growth on unusual metabolites exchanged in complex microbial communities are often much less well understood, resulting in missing or lower confidence functional annotations in newly sequenced genomes. 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.

Introduction

The process of annotating genomes has existed since the first sequenced genomes. Initially, bacterial genomes were annotated primarily through manual curation of different functional genes by teams of experts. Today, 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. Many draft genomes and metagenome bins are often run through a single annotation pipeline where genome annotations are inherited from previous genome annotations. These annotation tools each have their own distinct strengths and weaknesses, often trained on particular aspects of annotation or particular organisms. There is a trade-off between general annotation tools which excel at identifying genes in core metabolic processes, and specialist annotation tools focusing on a particular set of genes or enzyme substrates (such as

transporters). The range of different tools can be overwhelming, and the burden of learning the intricacies of multiple annotation pipelines can lead the typical researcher to choose only a single tool for annotation of an entire genome. Even when multiple annotation tools are used, incorporating the different pieces of information in a cohesive manner remains a major barrier. Individual metabolic annotation tools often return annotations for different subsets of genes, however, offering the potential to greatly increase the completeness of metabolic annotations by combining the outputs of multiple tools.

“Genome-scale” metabolic models implicitly assume complete and accurate genome annotation. However, only 50-70% of genes in a typical prokaryotic genome are annotated (1), although that is estimated to include a large (and potentially growing) fraction of misannotations (Schnoes 2009), and more than 30% of the unannotated genes are estimated to have metabolic functions as well (Ellens 2017), leaving a significant gap in our understanding of the underlying metabolic processes. The most reliably annotated genes are often those involved in core metabolic processes. Modern metabolic modeling efforts, however, are moving beyond studying core metabolic pathways in a single organism towards multi-species models, real-world communities and ecosystems (Henry et al. 2016). Additionally, incorporation of complex ‘omics and metabolite data is becoming increasingly common, and robust genome annotations are therefore necessary to aggregate these different data streams.

In Flux Balance Analysis (FBA), 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 (Thiele et al. 2010). 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 (2-5). Henry et al. 2010 showed that in a broad collection of 130 genome-scale metabolic models added to the ModelSEED database, on average 56 additional gap filled reactions were needed for each model to produce biomass in simple defined nutrient media (6). Even after those additions, however, one-third of the reactions in each model were still inactive, meaning that there were enough holes left in the network to preclude metabolic flux through those reactions (6,7). In addition, the number of reactions that can partake in a gap filling solution is vast (3,270 in the case of E. coli (8)), and the sets of reactions generated by different gap filling algorithms may have little or no overlap with each other (8). 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 (9) 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 (10), KEGG (11), and RefSeq annotations captured in BioCyc (12)) to evaluate annotation coverage and produce a more robust model. Reactions identified by all three methodologies were considered most reliable, however, this accounted for only 34 percent of the reactions. The overlap between annotations from different sources corresponded with an active set of reactions that represented the core

metabolic network of *C. beijerinckii* and illustrated the importance of annotation agreement. However, while this convergence of annotations is true for primary metabolic pathways, a comparison of annotations for secondary metabolic pathways is less clear.

In addition to modeling the sets of reactions *in vivo*, accurate metabolic models also rely on accurate determination of substrate transport between the bacterium and its environment. Transporter annotations have been rarely 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 bacteria *Streptococcus thermophilus* has an unusual growth phenotype in that it grows much better on lactose than on glucose (13), even though it does so by importing lactose, hydrolyzing it to glucose and galactose, and then metabolizing only the glucose, with the galactose secreted back out of the cell. The cause for its growth deficit when fed on glucose directly is that it lacks the usual glucose phosphotransferase system used by many bacteria, and instead it has a very efficient lactose import mechanism that makes it well adapted to grow in milk (13). Better prediction tools such as TransportDB's Transporter Automatic Annotation Pipeline (TransAAP (14)) now allow us 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 individual tools in genome annotation and their overlap with each other. Using 27 bacterial reference genomes from BioCyc (12), we evaluated how many genes, EC numbers, and gene-reaction links were unique or shared with the findings of other tools. We also undertook a study of how transporter annotations were handled between RAST (10), KEGG (11), and TransportDB (14).

Materials and Methods

Reference Genomes

We focused on a total of 27 genomes of BioCyc Tier 1 & Tier 2 bacteria (12). 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) and pathogens (e.g. *Mycobacterium tuberculosis*), an obligate insect endosymbiont (*Candidatus* Evansia muelleri), an unusual manganese oxidizing bacterium (*Aurantimonas manganoxydans* SI85-9A1), and an important marine primary producer (*Synechococcus elongatus* PCC 7942).

Genbank files were downloaded from NCBI (accessions list in Table 1), and standardized to remove all functional annotation, retaining only the original open reading frames and locus tag/protein identification information.

Table 1. Reference genomes used in this study

Genome Name	Biocyc ID	Phylum	NCBI Accessions	Proteins
Mycobacterium tuberculosis CDC1551	MTBCDC1551	Actinobacteria	AE000516	4189
Mycobacterium tuberculosis H37Rv	MTBH37RV	Actinobacteria	AL123456	4018
Streptomyces coelicolor A3(2)	SCO	Actinobacteria	NC_003888, NC_003903, NC_003904	8152
Bacteroides thetaiotaomicron VPI-5482	BTHE	Bacteroidetes	AE015928, AY171301	4825
Candidatus Cardinium hertigii[b]	CBTQ1	Bacteroidetes	HG422566, CBQZ010000001-CBQZ010000011	739
Synechococcus elongatus PCC 7942	SYNEL	Cyanobacteria	CP000100, CP000101	2661
Listeria monocytogenes 10403S	10403S_RAST	Firmicutes	CP002002	2814
Bacillus anthracis Ames	ANTHRA	Firmicutes	NC_003997, AE017335, AE017336	5602
Bacillus subtilis 168	BSUB	Firmicutes	AL009126	4185
Clostridium saccharoperbutylacetonicum ATCC 27021	CLOSSAC	Firmicutes	CP004121, CP004122	5821
Eubacterium rectale ATCC 33656	EREC	Firmicutes	CP001107	3626
Peptoclostridium difficile 630	PDIF272563	Firmicutes	AM180355, AM180356	3809
Agrobacterium fabrum C58	AGRO	Proteobacteria	AE008687, AE008688, AE008689, AE008690	5402
Aurantimonas manganoxydans SI85-9A1	AURANTIMONAS	Proteobacteria	AAPJ01000001-AAPJ01000035	3650
Caulobacter crescentus CB15	CAULO	Proteobacteria	AE005673	3737
Caulobacter crescentus NA1000	CAULONA1000	Proteobacteria	CP001340	3885
Escherichia coli CFT073	ECOL199310	Proteobacteria	AE014075	5379
Escherichia coli K-12 substr. W3110	ECOL316407	Proteobacteria	NC_007779	4410
Escherichia coli B str. REL606	ECOL413997	Proteobacteria	CP000819	4209
Escherichia coli K-12 substr. MG1655[a]	ECOLI	Proteobacteria	U00096	4140
Escherichia coli O157:H7 str. EDL933	ECO0157	Proteobacteria	AE005174, AF074613	5449
Candidatus Evansia muelleri[b]	EVA	Proteobacteria	LM655252	330
Helicobacter pylori 26695	HPY	Proteobacteria	CP003904	1594
Methylosinus trichosporium OB3b	MOB3B	Proteobacteria	NZ_ADVE02000001-NZ_ADVE02000003	4344
Candidatus Portiera aleyrodidarum BT-QVLC[b]	PABTQVLC	Proteobacteria	CP003867	280
Shigella flexneri 2a str. 2457T	SHIGELLA	Proteobacteria	AE014073	4068
Vibrio cholerae O1 biovar El Tor str. N16961	VCHO	Proteobacteria	AE003852, AE003853	3828

a: Tier 1 Pathway Genome Database (EcoCyc)

b: Endosymbiont with reduced genome

Annotation Tools

RAST (Rapid Annotation Subsystem Technology, (10)) is an open-source web server for genome annotation, using a "Highest Confidence First" 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, as supported by comparative and other forms of evidence. We used the NMPDR website (rast.nmpdr.org) to generate genome-wide annotations for our 27 reference genomes, and parsed any EC numbers from the functional annotation.

KEGG (Kyoto Encyclopedia of Genes and Genomes, (11)) is a collection of genome and pathway databases for systems biology. We used the KAAS (KEGG Automatic Annotation Server, www.genome.jp/tools/kaas, (15)) 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 (http://www.genome.jp/kegg-bin/get_htext?ko01000.keg).

EFICAz (16) 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 as low as 40%. We used a local install of EFICAz2.5 to generate EC number predictions for our 27 reference genomes.

BRENDA (17) 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 BRENDA database of enzyme reference sequences.

Transporter annotations

Where available, pregenerated transporter annotations were downloaded from the TransportDB 2.0 database (14). 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 (http://www.membranetransport.org/transportDB2/TransAAP_login.html). We also retrieved transporter annotations from RAST and KEGG. For RAST, annotations with a corresponding feature ID were assigned a SEED subsystem (http://rast.nmpdr.org/seedviewer.cgi) and filtered for the subsystem for “Membrane Transport”. Then, we linked to the locus tag of the matching feature ID. For KEGG, we mapped KO numbers to transporter annotations using a mapping table provided by the KEGG BRITE Database (http://www.genome.jp/kegg-bin/get_htext?ko02000.keg). Substrate names were ranked from most to least specific.

Substrates that can be incorporated as a transport reaction in a metabolic model are ranked 1 and 2. Broader substrate classes that could be used for gap filling or interpretation of transcriptomics data are ranked 3 and 4, and annotated transporters without substrate prediction are ranked 5. The full table of substrate names and ranking can be found in Supplementary Materials.

Table 2. Examples of substrate annotation ranking, from most specific (rank 1) to least specific (no substrate, rank 5)

Rank	Substrate	Examples
1	Metabolite that can be incorporated as a transport reaction in a metabolic model	● Fe ● lysine
2	Substrate(s) that map to a small number of possible transport reactions	● Mg/Co/Ni ● aromatic amino acid
3	Broader substrate classes not directly usable to construct a metabolic network	● dipeptide ● sugar
4	Very broad class of substrates	● multidrug efflux ● protein
5	No substrate annotated	

Results

GENE-EC number annotations

In total, the four annotation tools produced 47,447 unique 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 can differ drastically (Figure 1). Each tool produced on average between 23% (EFICAz) and 48% (BRENDA) unique gene-EC annotations that were not predicted by any of the other tools. Overall, fewer than a quarter of all gene-EC annotations are agreed on by at least 3 tools.

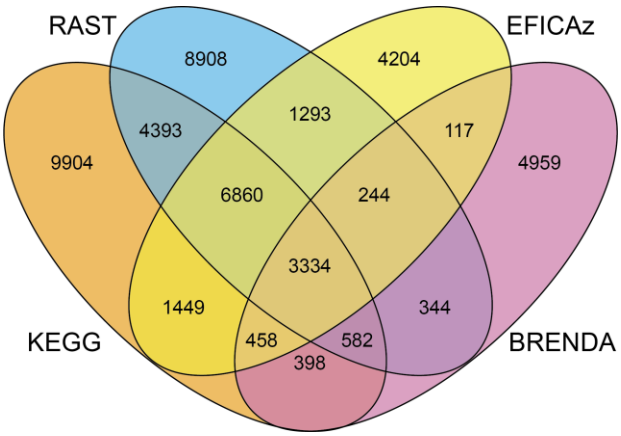

Figure 1. Large differences exist between the sets of Gene-EC annotations generated by the four annotation tools across the 27 reference genomes.

Table 3 illustrates the number of gene-EC agreements among pairs of annotation tools. The denominator is the size of the intersection in gene annotation listings between the pair of tools. In other words, out of the genes that are given some annotation (i.e. genes in the intersections in Figure 3B), what fraction is overlapping. If one tool does not have an annotation (or an incomplete classification consisting of less than 4 numbers), the comparison is not made. We are also not requiring that tools produce identical gene-EC numbers if one or both of the tools produce multiple gene-EC numbers per annotation (a more stringent criteria), only that there is at least one agreement between the two. This is because some tools may be more sensitive to picking up secondary enzymatic functions.

Table 3. 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	3,777/6,748 (56.0%)
RAST-EFICAz	12,977/15,694 (82.7%)
RAST-BRENDA	3,907/6,288 (62.1%)
EFICAz-BRENDA	3,902/5,601 (69.7%)

The findings show that combining BRENDA with other tools produces fewer agreements--this dovetails with the finding that BRENDA tends to produce more unique annotations. On the other hand, a pairing with EFICAz shows more agreements. KEGG and RAST have much bigger sets

of annotations but they tend to produce low disagreement with other tools which makes them useful for verification but less informative about those not annotated by either of them.

The EcoCyc database (18) is an extensively hand-curated and continuously updated database summarizing all the experimentally determined enzymatic functions in *Escherichia coli* K-12 substr. MG1655, the single best studied model organism in the history of modern biology. We can use the gene-EC numbers annotated in EcoCyc as a gold standard to evaluate how well the automated annotation tools are able to annotate the enzymes in *E. coli* K-12. Figure 2 shows how the set of gene-EC annotations generated by KEGG and RAST compare against EcoCyc.

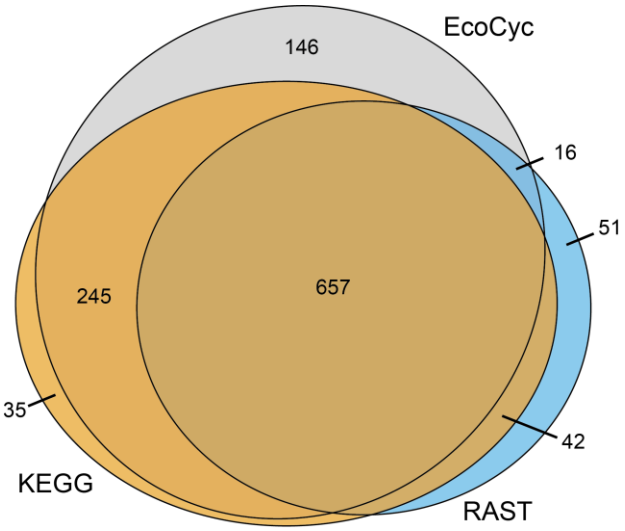

Figure 2. Gene-EC annotations produced by KEGG and RAST for *E. coli* K-12, compared to the EcoCyc gold standard. The sets and intersections are drawn proportionally to the number of annotations in each.

EC number annotations

While gene-EC links give a handle on the quality of the individual annotations, using the EC numbers alone gives a handle on the of the quality and completeness of the metabolic network. The presence of isozymes in which similar reactions are catalyzed by different enzymes also adds an important dimension to the analysis.

Figure 3A shows the number of EC numbers annotated by each of the tools across the 27 reference genomes, which reflects the size of the metabolic network reconstruction (average 868 EC numbers per genome). Figure 3B shows the number of genes in all of the reference genomes annotated with one or more EC numbers by each of the tools, which reflects the overall genome annotation coverage (average 1361 genes per genome, or 34% of the genes).

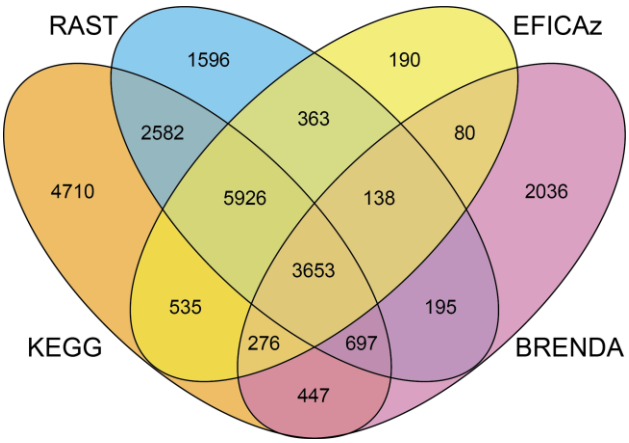

Figure 3. Combining Annotation Tools Improves Genome annotation and Metabolic Network Coverage. A: Total unique EC numbers annotated. B: Total genes annotated with EC numbers

Note that each tool adds a significant number of reactions to the metabolic network model, and each tool 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 tend to belong to well studied core metabolic pathways. Out of the 79 EC numbers on which all four tools agree in at least half of the genomes, more than three quarters (61/79) are 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 are involved in biosynthesis or degradation of fatty acids, lipids, aromatics compounds, or secondary metabolites.

The disjoint sets of annotations produced by the different tools provide us with an opportunity to trade off confidence in the annotations versus coverage. If higher confidence is required, we can focus solely on the subset of annotations that is agreed upon by multiple tools. Conversely, if the lack of genome coverage or metabolic network coverage is considered a problem, we can use the union of multiple tools to achieve a wider annotation.

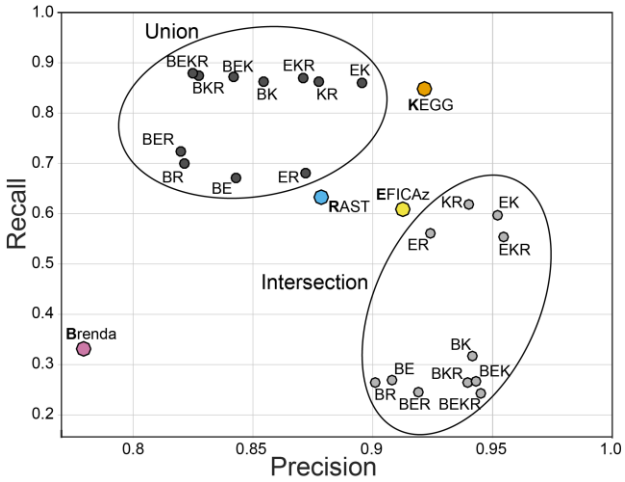

Figure 4. Precision vs Recall of EC numbers for different combinations of tools on EcoCyc. Individual tools are denoted by B,E,K, or R for BRENDA, EFICAz, KEGG, and RAST, respectively.

Choosing *E. coli* K-12 as the "gold standard", we took the EC numbers annotated by each tool and performed the union and intersection of each of the combinations. These combinations included all pairs, triplets as well as the list of ECs from the union and intersection of all 4 tools combined. We then compared each of these EC lists to the 1064 EC numbers from EcoCyc. The number of true positives, false positives, and false negatives were calculated for each combination. True positives (TP) correspond to EC numbers predicted by the annotation tools, and present in EcoCyc. False positives (FP) are EC numbers annotated by the tools, but not found in EcoCyc. False negatives (FN) are those EC number that occur in EcoCyc, but were not predicted by our annotation tools. Precision is defined as the (TP)/(TP+FP) and recall as (TP)/(TP+FN).

Figure 4 shows a plot of precision versus recall for all the different combinations of tools. Three groupings of points stand out from this graph. First, all the combinations that contain some union of the tools have high precision and high recall. Second, the combinations that are a result of intersections of EC lists are bunched in an area of very high precision but much lower recall. In between these two groups lie the single tool results. One can see that precision and recall are lowest for BRENDA alone; KEGG in this case has the best precision and recall of the four alone. When any set of tools intersects with BRENDA, precision remains about the same but recall drops dramatically (e.g. comparing ER and BER). On the other hand, a union with KEGG increases recall. In some cases, the combination of tools makes the precision and/or recall worse.

Using EcoCyc as the Gold Standard for *E. coli* K-12, the different tools achieve a Precision of 78% (BRENDA) to 92% (KEGG), and a Recall of 33% (BRENDA) to 85% (KEGG) when comparing their EC numbers. Note that even for *E. coli* K-12, which we expect to be a best-case situation, there are significant differences in the annotations produced by the different tools, and each tools is only able to cover a subset of the known enzymes in EcoCyc.

The annotation tools show much more agreement on E. coli than on more remote lineages such as Actinomycetes, Bacteroidetes, or Clostridia. For *E. Coli* K-12, 60% of EC numbers are agreed on by 3 or more tools, while 28% EC numbers come from only a single tool. In contrast, for *P. difficile* 630, only 33% of EC numbers are agreed on by 3 or more tools, and 48% of EC numbers come from only a single tool.

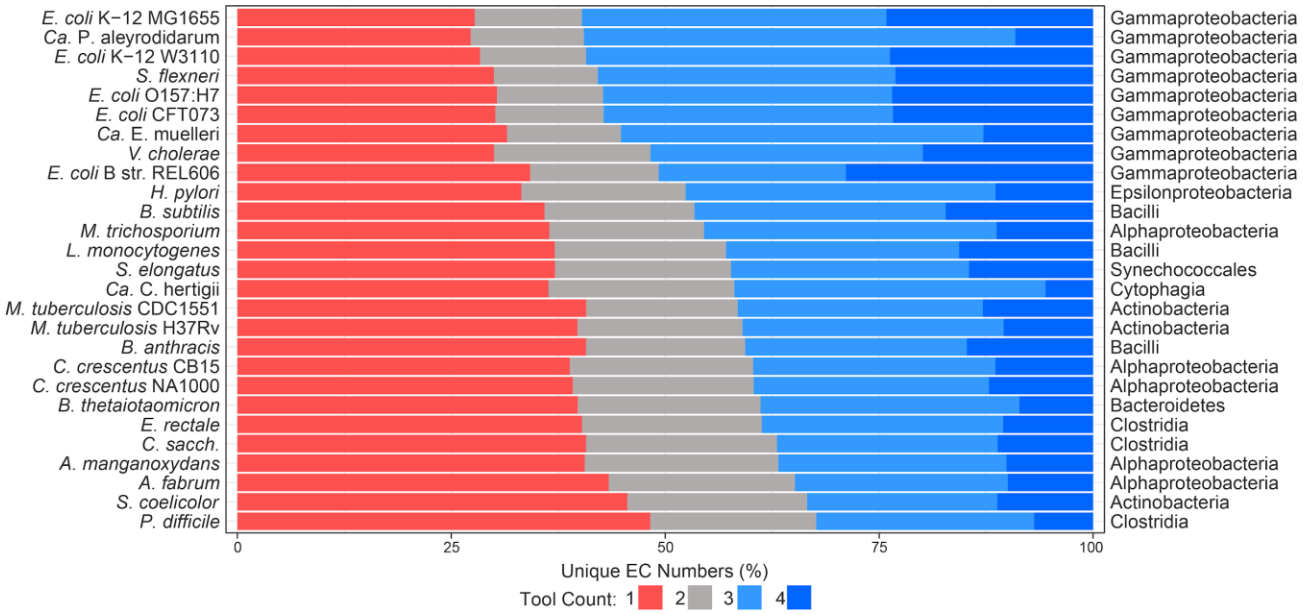

Figure 5. The 27 reference genomes were sorted with respect to the fraction of EC numbers that were predicted by 3 or more tools (blue bars). The top of the list is dominated by model organisms such as *E. coli*, *B. subtilis*, and closely related organisms. As we move farther away from such well-studied model organisms, the fraction of unique EC numbers predicted only by a single tool (red bars) increases, at the expense of those predicted by multiple tools.

Transporter Annotations

Both RAST and KEGG yield surprisingly few transporter annotations, with an average of 114 and 204 transporter predictions per genome respectively. In addition, many of the annotated transporters lack substrate predictions (52% of transporter annotations in RAST, 25% in KEGG) or have ambiguous substrate predictions (ranks 3-4; 20% in RAST, 28% in KEGG), while less than half have substrate predictions that are sufficiently detailed that they could be incorporated in a metabolic model (ranks 1-2; 28% in RAST, 48% in KEGG). 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).

Transporter annotations by the different tools show remarkable little overlap (Figure 6). Out of the more than 15,000 genes annotated as transporters (regardless of substrate prediction), the three tools only agree on 2.8% (423/15161). Out of those, only 130 genes are annotated by all three tools with a specific substrate prediction (ranks 1-2)

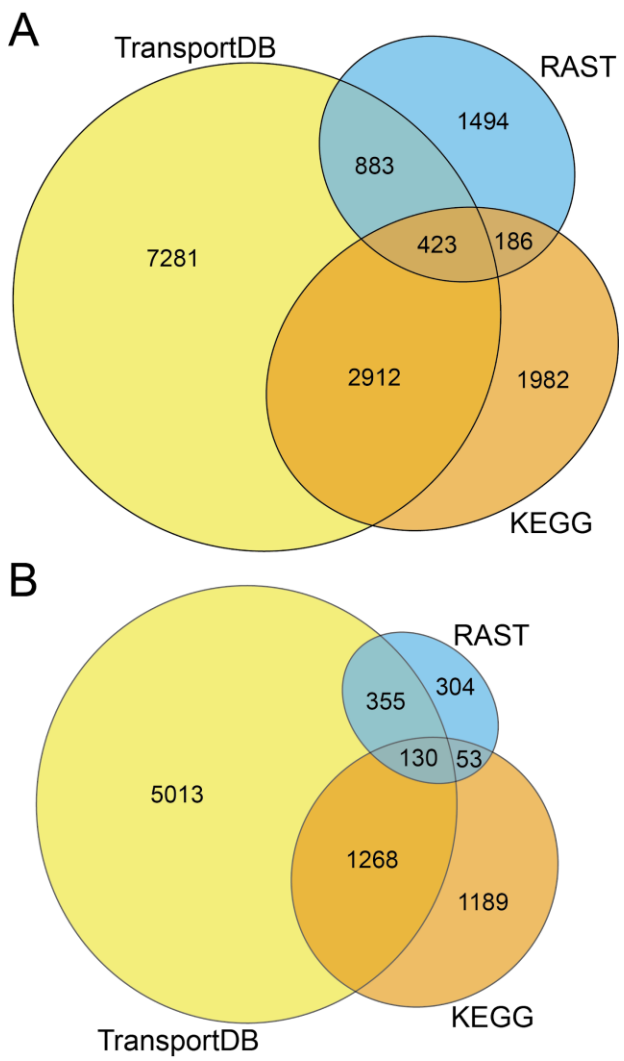

Figure 6. A: Total number of genes annotated as transporters, regardless of substrate. B: Transporter annotations with substrates predictions specific enough to be included in metabolic models (rank 1 or 2).

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 perfectly identical (for example, one transporter was annotated as “leucine/valine”, “leucine”, and “branched-chain amino acid” by TransportDB, RAST and KEGG respectively).

Discussion

RAST and KEGG are the most widely used tools for metabolic network reconstruction. However, they do not necessarily produce identical annotations. In our analysis, KEGG produced the best Precision and Recall on *E. coli* K-12. In general, KEGG produces a larger number of unique EC numbers, which could indicate more over-prediction, or more

comprehensive pathway coverage. Note that both also generate many reactions without official EC numbers.

EFICAz produces the least number of unique EC numbers, but can be used in combination with RAST or KEGG to highlight high confidence annotations. EFICAz also produced incomplete “three-digit” EC number annotations (*e.g.* 1.2.3.-) which may be used for hole filling.

BLAST against the BRENDA database of reference enzymes produced the smallest number of annotations, but a high fraction of unique EC numbers. 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, and six are EC numbers that have been assigned in 2000 or newer, and may not have been incorporated into the predictions by the other annotation tools yet.

This analysis has led us to make a number of recommendations when approaching an annotation issue. We believe that one should not use just a single annotation tool unless one is only interested in core metabolism where tools often agree. In addition, one can trade off confidence in predictions versus greater coverage by the intersection or union of multiple annotation tools. Even tools like EFICAz that produce fewer unique annotations provide higher confidence in them, and also provide partial EC numbers for hole filling. 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 effort to bring together annotation from multiple sources. More tool development is needed to merge annotations beyond simple EC numbers.

Recommendations

1. Do not just use a single annotation tool unless you are only interested in core metabolism.
2. Trade off confidence versus coverage by looking at the intersection or union of multiple annotation tools.
3. EFICAz can be used to identify higher confidence annotations, or partial EC numbers for hole filling
4. BLASTing against a database of reference sequences such as the BRENDA database 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.
5. More tool development is needed to allow merging of annotations beyond simple EC numbers.

Funding

This work was supported by the Department of Energy through the Genome Sciences 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.

References

1. Hanson, A.D, Pribat, A., Waller, J.C. and de Crécy-Lagard, V. (2009) 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list--and how to find it. *Biochem J.*, 425(1), 1-11.

- Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 2009 Dec;5(12):e1000605.
- 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 Nov 16;45(20):11495-11514.
- Henry CS, Bernstein HC, Weisenhorn P, Taylor RG, Taylor RC, Lee J-Y, Zucker J, Song H-S (2016). Microbial Community Metabolic Modeling: A Community Data-Driven Network Reconstruction. J Cell Physio., 231(11): 2339-45.
- Thiele I, Palsson, BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature Protoc., 5:93–121.

2. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO. (2006) Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A*. 103(46):17480-4.
3. Satish Kumar V, Dasika MS, Maranas CD. (2007) Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics.* 8:212.
4. Kumar VS, Maranas CD. (2009) GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. *PLoS Comput Biol.* 5(3):e1000308.
5. Benedict MN, Mundy MB, Henry CS, Chia N, Price ND. (2014) Likelihood-based gene annotations for gap filling and quality assessment in genome-scale metabolic models. *PLoS Comput Biol*. 10(10):e1003882.
6. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol*. 28(9):977-82
7. Ponce-de-Leon M, Calle-Espinosa J, Peretó J, Montero F. (2015) Consistency Analysis of Genome-Scale Models of Bacterial Metabolism: A Metamodel Approach. *PLoS One.* 10(12):e0143626.
8. Krumholz EW, Libourel IG. (2015) Sequence-based Network Completion Reveals the Integrality of Missing Reactions in Metabolic Networks. *J Biol Chem*. 2015 290(31):19197-207.
9. Milne, C.B., Eddy JA, Raju, R., Ardekani, S., Kim, P.J., Senger, R.S., Jin, Y.S., Blaschek, H.P. and Price, N.D. (2011). Metabolic network reconstruction and genome-scale model of butanol-producing strain Clostridium beijerinckii NCIMB 8052. *BMC Syst Biol.*, 5, 130.
10. Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A.R., Xia, F. and Stevens, R. (2014)

The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.*, 42(Database issue), D206-214.
11. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44(D1):D457-62.
12. Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S. and Karp, P.D. (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, 44(D1), D471-D480.
13. Poolman B. (1993) Energy transduction in lactic acid bacteria. *FEMS Microbiol Rev.* 12(1-3):125-47
14. Elbourne, L.D.H., Tetu, S.G., Hassan, K.A., Paulsen, I.T. (2017) TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic Acids Res*. 45, D320–D324.
15. Moriya, Y., Itoh, M, Okuda, S., Yoshizawa, A.C. and Kanehisa, M. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, 35(Web Server issue), W182-W185.
16. Kumar, N. and Skolnick, J. (2012) EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes. *Bioinformatics*, 28(20), 2687-2688.
17. Placzek, S., Schomburg, I., Chang, A., Jeske, L., Ulbrich, M., Tillack, J. and Schomburg, D. (2017) BRENDA in 2017: new perspectives and new tools. *Nucleic Acids Res.*, 35(D1), D380-388.
18. Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martínez C, Caspi R, Fulcher C, Gama-Castro S, Kothari A, Krummenacker M, Latendresse M, Muñiz-Rascado L, Ong Q, Paley S, Peralta-Gil M, Subhraveti P, Velázquez-Ramírez DA, Weaver D, Collado-Vides J, Paulsen I, Karp PD. (2017) The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Res.* 45(D1):D543-D550.