RAVEN 2.0: a versatile platform for metabolic network reconstruction and a case study on *Streptomyces coelicolor*

Hao Wang[1,¶], Simonas Marcišauskas[1,¶], Benjamín J Sánchez[1], Iván Domenzain[1], Daniel Hermansson[1], Rasmus Agren[2], Jens Nielsen[1,3], Eduard J Kerkhoven[1,*]

[1] Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, SE412 96 Gothenburg, Sweden

[2] Department of Biology and Biological Engineering, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Chalmers University of Technology, Kemivägen 10, SE412 96 Gothenburg, Sweden

[3] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, DK2800 Lyngby, Denmark

* Corresponding author.

Email: eduardk@chalmers.se (EJK)

¶ These authors contributed equally to this work.

Abstract

RAVEN is a commonly used MATLAB toolbox for genome-scale metabolic model (GEM) reconstruction, curation and constraint-based modelling and simulation. Here we report RAVEN Toolbox 2.0 with major enhancements, including: (i) de novo reconstruction of GEMs based on the MetaCyc pathway database; (ii) a redesigned KEGG-based reconstruction pipeline; (iii) convergence of reconstructions from various sources; (iv) improved performance, usability, and compatibility with COBRA Toolbox. Capabilities of RAVEN 2.0 were demonstrated through de novo reconstruction of GEMs for the antibiotic-producing bacterium *Streptomyces coelicolor*. Comparison of the automated de novo reconstruction with iMK1208, a previously published high-quality *S. coelicolor* GEM, demonstrated that RAVEN 2.0 captures most of the manually curated model. The generated de novo reconstruction was subsequently used to curate iMK1208 resulting in Sco4, the most comprehensive GEM of *S. coelicolor*, with increased coverage of both primary and secondary metabolism. This increased coverage allowed the use of *Sco4* to predict novel genome editing targets for optimized secondary metabolites production. As such, we demonstrate how RAVEN 2.0 can not only be used for de novo GEM reconstruction, but also for curating existing models based on up-to-date databases. Both RAVEN 2.0 and *Sco4* are deposited on GitHub to facilitate usage and further development by the community.

Author summary

Cellular metabolism is a large and complex network. Hence, investigations of metabolic networks are aided by *in silico* modelling and simulations. Metabolic networks can be derived from whole-genome sequences, through identifying what enzymes are present and connect these to formalized chemical reactions. To facilitate the reconstruction of genome-scale models of metabolism (GEMs), we have developed RAVEN 2.0. This versatile toolbox can reconstruct GEMs fast, through either metabolic pathway databases KEGG and MetaCyc, or from homology with an existing GEM. We demonstrate RAVEN functionality through generation of a metabolic model of *Streptomyces coelicolor*, an antibiotic-producing bacterium. Comparison of this *de novo* generated GEM with a previously manually curated model demonstrates that RAVEN captures most of the previous model, and we subsequently reconstructed an updated model of *S. coelicolor*: *Sco4*. Following, we used *Sco4* to predict promising targets for genetic engineering, which can be used to increase antibiotic production.

Introduction

Genome-scale metabolic models (GEMs) are comprehensive *in silico* representations of the complete set of metabolic reactions that take place in a cell [1]. GEMs can be used to understand and predict how organisms react to variations on genetic and environmental parameters [2]. Recent studies demonstrated the extensive applications of GEMs in discovering novel metabolic engineering

strategies [3]; studying microbial communities [4]; finding biomarkers for human diseases and personalized and precision medicines [5,6]; and improving antibiotic production [7]. With the increasing ease of obtaining whole-genome sequences, which are the basis for reconstructing GEMs, the challenge remains to translate this knowledge to high-quality GEMs [8].

To meet the increasing demand of metabolic network modelling, the original RAVEN (Reconstruction, Analysis and Visualization of Metabolic Networks) toolbox was developed to facilitate GEM reconstruction, curation, and simulation [9]. In addition to facilitating the analysis and visualization of existing GEMs, RAVEN particularly aimed to assist semi-automated draft model reconstruction, utilizing existing template GEMs and the KEGG database [10]. Since publication, RAVEN has been used in GEMs reconstruction for a wide variety of organisms, ranging from bacteria [11], archaea [12] to human gut microbiome [13], eukaryotic microalgae [14], parasites [15,16], and fungi [17], as well as various human tissues [18,19]. As such, the RAVEN toolbox has functioned as one of the two major MATLAB-based packages for constraint-based metabolic modelling, together with the COBRA Toolbox [20,21].

Here, we present RAVEN 2.0 with greatly enhanced reconstruction capabilities, together with additional new features (Fig 1). A prominent enhancement of RAVEN 2.0 is the use of the MetaCyc database in assisting draft model reconstruction. MetaCyc is a pathway database that instead of relying on sequence-based annotations (similar to KEGG), collects only experimentally verified pathways with curated reversibility information and mass-balanced reactions [22]. RAVEN 2.0 can leverage this high-quality database to enhance the GEM reconstruction process. We demonstrate the new and improved capabilities of RAVEN 2.0 through reconstruction of a GEM for *Streptomyces coelicolor*.

Streptomyces coelicolor is a representative species of soil-dwelling, filamentous and gram-positive actinobacterium harbouring enriched secondary metabolite biosynthesis gene clusters [23,24]. As a well-known pharmaceutical and bioactive compound producer, *S. coelicolor* has been exploited for antibiotic and secondary metabolite production [25]. The first published GEM for *S. coelicolor*, iIB711 [26], was improved through an iterative process, resulting in the GEMs iMA789 [27] and iMK1208 [28]. The most recent GEM, iMK1208, is a high-quality model that includes 1208 genes and 1643 reactions, and was successfully used to predict metabolic engineering targets for increased production of actinorhodin [28].

Here, we demonstrate how the new functions of RAVEN 2.0 can be used for *de novo* reconstruction of a *S. coelicolor* GEM, using comparison to the existing high-quality model iMK1208 as benchmark. The use of three distinct *de novo* reconstruction approaches enabled capturing most of the existing model, while complementary reactions found through the *de novo* reconstructions gave the opportunity to improve the existing model. After manual curation, we included 402 new reactions into the GEM, with 320 newly associated enzyme-coding genes, including a variety of biosynthetic

pathways for known secondary metabolites (e.g. 2-methylisoborneol, albaflavenone, desferrioxamine, geosmin, hopanoid and flaviolin dimer). The updated *S. coelicolor* GEM is released as *Sco4*, which can be used as an upgraded platform for future systems biology research on *S. coelicolor* and related species.

Results

Genome-scale reconstruction and curation with RAVEN 2.0

RAVEN Toolbox 2.0 aims to provide a versatile and efficient platform for metabolic network reconstruction and curation (Fig 1). In comparison to other solutions for GEM reconstruction (Table 1), the strength of RAVEN 2.0 is its ability of semi-automated reconstruction based on published models, KEGG and MetaCyc databases, integrating knowledge from diverse sources. A brief overview of RAVEN capabilities is given here, while more technical details are stated in Material & Methods, and detailed documentation is provided for individual functions in the RAVEN package.

RAVEN supports two distinct approaches to initiate GEM reconstruction for an organism of interest: (i) based on protein homology to an existing template model, or (ii) *de novo* using reaction databases. The first approach requires a high-quality GEM of a phylogenetically closely related organism, and the functions `getBlast` and `getModelFromHomology` are used to infer homology using bidirectional BLASTP and build a subsequent draft model. Alternatively, *de novo* reconstruction can be based on two databases: KEGG and MetaCyc. For KEGG-based reconstruction, the user can deploy `getKEGGModelForOrganism` to either rely on KEGG-supplied annotations—KEGG currently includes over 5000 genomes—or query its protein sequences for similarity to HMMs that are trained on genes annotated in KEGG. MetaCyc-based reconstruction can be initiated with `getMetaCycModelForOrganism` that queries protein sequences with BLASTP for homology to enzymes curated in MetaCyc, while `addSpontaneous` retrieves relevant non-enzyme associatedreactions from MetaCyc.

Regardless of which (combination of) approach(es) is followed, a draft model is obtained, that requires further curation to result in a high-quality reconstruction suitable for simulating flux distributions. Various RAVEN functions aid in this process, including `gapReport` that runs a gap analysis and reports e.g. dead-end reactions, unconnected subnetworks, metabolites that can be produced or consumed without in or output from the model (indicative of unbalanced reactions). While RAVEN is distributed with a gap-filling algorithm `gapFill`, however, results from external gap-filling approaches can be readily incorporated. This and further manual curation is facilitated through functions such as `addRxnsGenesMets` that moves reactions from a template to a draft model, `changeGeneAssoc` and `standardizeGrRules` that curate gene associations and

`combineMetaCycKEGGModels` that can semi-automatically unify draft models reconstructed from different databases.

In addition to model generation, RAVEN 2.0 includes basic simulation capabilities including flux balance analysis (FBA), random sampling of the solution space [29] and flux scanning with enforced objective function (FSEOF) [30]. Models can be handled in various file-formats, including the community standard SBML L3V1 FBCv2, compatible with COBRA Toolbox [20]; a user-friendly Excel format; flat-text and YAML formats to facilitate tracking changes between model versions during curation. As a MATLAB package, RAVEN 2.0 provides one unified environment supporting both model reconstruction and simulation analysis, while allowing scripting for more flexible operations.

Draft model reconstruction for *S. coelicolor*

The enriched capabilities of RAVEN 2.0 were evaluated by *de novo* generation of GEMs for *S. coelicolor* using three distinct approaches, as described in Material & Methods (Fig 2). Cross-comparison of genes from *de novo* reconstructions and the published *S. coelicolor* GEM iMK1208 indicated that the three *de novo* approaches are complementary and comprehensive, combined covering 88 % of the genes included in iMK1208 (Fig 3). The existing model contained 146 genes that were not annotated by any of the automated approaches, signifying the valuable manual curation that has gone into previous GEMs of *S. coelicolor*. Nonetheless, matching of metabolites across models through their KEGG identifiers further supported that most of the previous GEM is captured by the three *de novo* reconstructions, while each approach has their unique contribution (Fig 3).

The three draft reconstructions were consecutively merged to result in a combined draft reconstruction (S1 Data), containing 2605 reactions, of which 958 and 1104 reactions were uniquely from MetaCyc- and KEGG-based reconstructions, respectively (Fig 2). While MetaCyc-based reconstruction annotated more genes than KEGG-based reconstructions (Fig 3), the number of unique reactions by MetaCyc is slightly lower than by KEGG, indicating that KEGG based reconstruction is more likely to assign genes to multiple reactions. Of the 789 reactions from the existing high-quality model that could be mapped to either MetaCyc or KEGG reactions, 733 (92.9%) were included in the combined draft model (S1 Table).

Further development of the obtained model

The combined *de novo* reconstruction has a larger number of reactions, metabolite and genes than the previous *S. coelicolor* GEM (Fig 2). While a larger metabolic network does not necessarily imply a *better* network, we took advantage of the increased coverage of the *de novo* reconstruction by using it to curate iMK1208, while retaining the valuable contributions from earlier GEMs. The culminating

model is called *Sco4*, the fourth major release of *S. coelicolor* GEM. Through manual curation, a total of 398 metabolic reactions were selected from the combined model to expand the stoichiometric network of the previous GEM (S3 Table). These new reactions cover diverse subsystems including both primary and secondary metabolism (Fig 4A), and displayed close association with existing metabolites in the previous GEM (Fig 4B). Despite both MetaCyc- and KEGG-based reconstructions contributing roughly equally, MetaCyc-unique reactions are more involved in energy and secondary metabolism, while KEGG-unique reactions are more related to amino acid metabolism and degradation pathways (Fig 4C). The *de novo* reconstructions annotated genes to 11 reactions that had no gene association in the previous GEM (S4 Table). Together with 34 spontaneous reactions and 10 transport reactions identified by the MetaCyc reconstruction functions (S5 and S6 Tables), the resulting *Sco4* model contains 2304 reactions, 1927 metabolites and 1522 genes (Fig 2).

The process of model curation using *de novo* reconstructions furthermore identified erroneous annotations in the previous GEM. Four metabolites (involved in 12 reactions) were annotated with now invalid KEGG identifiers (Table 2), impeding matching with the KEGG-based reconstructions. However, mapping of the reactions via MetaCyc back to KEGG identified the metabolites, and in *Sco4* these metabolites are now annotated with valid KEGG identifiers. In a likewise manner, 14 other metabolites were curated by updating their invalid KEGG identifiers (S7 Table). While the identifiers used in the previous GEM were valid at the time, they were seemingly removed from the KEGG database in later versions. Unfortunately, neither KEGG nor MetaCyc provide changelogs for tracing such revisions.

Simulations and predictions with *Sco4*

The quality of *Sco4* was evaluated through various simulations. It displayed the same performance as iMK1208 in growth prediction on 64 different nutrient sources, with a consistent sensitivity of 90.6% (S8 Table). Experimentally measured growth rates in batch and chemostat cultivations were in good correlation with the growth rates predicted by *Sco4* (Fig 5A).

A recent large-scale mutagenesis study produced and analyzed 51,443 *S. coelicolor* mutants, where each mutant carried a single Tn5 transposition randomly inserted in the genome [31]. A total of 79 cold regions (> 1.9 kb), where no transposition insertions were identified, are indicative loci of potential essential genes, as insertion into these locations might result in a lethal phenotype. These cold regions coded for 132 genes, of which 65 are annotated to reactions in *Sco4* (S9 Table). Relevant is that cold regions of insertion are not proof of gene essentiality, as these cold regions can be caused by other mechanisms. Instead, it is more reasonable to assume that genes located *outside* cold regions are *not essential*; we therefore compared the non-essential gene sets. Simulation with *Sco4* indicates a specificity (or true negative rate) of 0.901, which is an increase over the 0.876 of the previous model (Fig 5B).

The *S. coelicolor* genome project revealed a dense array of secondary metabolite gene clusters both in the core and arms of the linear chromosome (Bentley et al. 2002), and extensive efforts have been made to elucidate these biosynthetic pathways (Van Keulen and Dyson, 2014). The previous GEM of *S. coelicolor* included only three of these pathways (i.e. actinorhodin, calcium-dependent antibiotic and undecylprodigiosin). Through our *de novo* reconstruction, we captured the advances that have since been made in elucidating additional pathways: *Sco4* describes the biosynthetic pathways of 6 more secondary metabolites (e.g. geosmin). These additional pathways were mainly contributed by the MetaCyc-based reconstruction (Fig 4B).

The expanded description of secondary metabolism was used to predict potential metabolic engineering targets for efficient antibiotic production in *S. coelicolor*. Flux scanning with enforced objective function (FSEOF) [30] was applied to all secondary metabolic pathways in *Sco4* and suggested overexpression targets were compared, with significant overlap between different classes of secondary metabolites (Fig 6, S10 Table). In addition, several targets were predicted to increase production of all modelled secondary metabolites. Three reactions, constituting the pathway from histidine to N-formimidoyl-L-glutamate, and catalyzed by SCO3070, SCO3073 and SCO4932, were commonly identified as potential targets (S10 Table).

Discussion

The RAVEN toolbox aims to assist constraint-based modeling with a focus on network reconstruction and curation. A growing number of biological databases have been incorporated for automated GEM reconstruction (Figure 1). The generation of tissue/cell type-specific models through task-driven model reconstruction (tINIT) has been incorporated to RAVEN 2.0 as built-in resource for human metabolic modeling [18,33]. RAVEN 2.0 was further expanded in this study by integrating the MetaCyc database, including experimentally elucidated pathways, chemically-balanced reactions, as well as associated enzyme sequences (21). This key enhancement brings new features toward high-quality reconstruction, such as inclusion of transport and spontaneous reactions (Table 1).

The performance of RAVEN 2.0 *de novo* reconstruction was demonstrated by the large overlap of reactions between the automatically obtained draft model of *S. coelicolor* and the manually curated iMK1208 model [28]. This indicates that *de novo* reconstruction with RAVEN 2.0 is an excellent starting point towards developing a high-quality model, while a combined *de novo* reconstruction can be produced within hours on a personal computer. We used the *de novo* reconstructions to curate the existing iMK1208 model, and the resulting *Sco4* model was expanded with numerous reactions, metabolites and genes, in part representing recent progress in studies on metabolism of *S. coelicolor* and related species (Figure 5). We have exploited this new information from biological databases to predict novel targets for metabolic engineering toward establishing *S. coelicolor* as a potent host for a

wide range of secondary metabolites (Figure 7). Therefore, RAVEN 2.0 can be used not only for *de novo* reconstruction but also model curation and continuous update, which would be necessary for a published GEM to synchronize with the incremental knowledge. We thus deposited the *Sco4* as open GitHub repositories for collaborative development with version control.

Material and Methods

RAVEN Toolbox development

The RAVEN Toolbox 1.0 was released as an open-source MATLAB-package [9], that has since seen minor updates and bugfixes. Since 2016, the development of RAVEN has been organized and tracked at a public GitHub repository. This repository provides a platform for the community, who are encouraged to report bugs, request new features and contribute to the development.

The RAVEN Toolbox is based on a defined model structure (S11 Table). Design choices dictate minor differences between COBRA and RAVEN structures, however, bi-directional model conversion is supported through `ravenCobraWrapper`. Through resolving previously conflicting function names, RAVEN 2.0 is now fully compatible with the COBRA Toolbox.

MetaCyc-based reconstruction module

Novel algorithms were developed to facilitate *de novo* GEM reconstruction by utilizing the MetaCyc database [22]. In this module, corresponding Matlab structures were generated from MetaCyc data files (version 21.0) that contained 3118 manually curated pathways with 13,689 metabolites and 15,309 reactions (Fig 7). A total of 17,394 enzymes are associated to these pathways and their protein sequences are included (protseq.fsa). Information from these structures is parsed by `getModelFromMetaCyc` to generate a model structure containing all metabolites, reactions and enzymes. This MetaCyc model can subsequently be used for *de novo* GEM reconstruction through the `getMetaCycModelForOrganism` function (Fig 7). A draft model is generated from MetaCyc enzymes (and associated reactions and metabolites) that show homology to the query protein sequences. Beneficial is that MetaCyc reactions are mass- and charge-balanced, while curated transport enzymes in MetaCyc allow inclusion of transport reactions into the draft model.

In addition, MetaCyc provides 515 reactions that may occur spontaneously. As such reactions have no enzyme association, they are excluded from sequence-based reconstruction and can turn into gaps in the generated models. By cataloguing spontaneous reactions in MetaCyc, the `addSpontaneousRxns` function can retrieve spontaneous reactions depending the presence of the relevant reactants in the draft model.

KEGG-based reconstruction module

In addition to MetaCyc-based GEM reconstruction, RAVEN 2.0 can utilize the KEGG database for *de novo* GEM reconstruction. The reconstruction algorithms were significantly enhanced in multiple aspects: the reformatted KEGG database in Matlab format is updated to version 82.0; and the pipeline to train KEGG Orthology (KO)-specific hidden Markov Models is expanded. Orthologous protein sequences, associated to particular KEGG Orthology (KO), are organised into non-redundant clusters with CD-HIT [34]. These clusters are used as input in m-sequence alignment with MAFFT [35], for increased accuracy and speed. The hidden Markov models (HMMs) are then trained for prokaryotic and eukaryotic species with various protein redundancy cut-offs (100%, 90% or 50%) using HMMER3 [36] and can now be automatically downloaded when running `getKEGGModelForOrganism`.

Combining of MetaCyc- and KEGG-based draft models

To capitalize on the complementary information from MetaCyc- and KEGG-based reconstructions, RAVEN 2.0 facilitates combining draft models from both approaches into one unified draft reconstruction (Fig 2). Prior to combining, reactions shared by MetaCyc- and KEGG-based reconstructions are mapped using MetaCyc-provided cross-references to their respective KEGG counterparts (S12 Table). Additional reactions are associated by `linkMetaCycKEGGRxns` through matching the metabolites, aided by cross-references between MetaCyc and KEGG identifiers (S13 Table). Subsequently, the `combineMetaCycKEGGModels` function thoroughly queries the two models for identical reactions, discarding the KEGG versions while keeping the corresponding MetaCyc reactions. In the combined model, MetaCyc naming convention is preferentially used such that unique metabolites and reactions from KEGG-based draft model are replaced with their MetaCyc equivalents whenever possible. The combined draft model works as a starting point for additional manual curation, to result in a high-quality reconstruction.

Miscellaneous improvements

RAVEN 2.0 contains a range of additional enhancements. Linear problems can be solved through either the Gurobi (Gurobi Optimization Inc., Houston, Texas) or MOSEK (MOSEK ApS, Copenhagen, Denmark) solvers. Various file formats are supported for import and export of models, including Microsoft Excel through Apache POI (The Apache Software Foundation, Wakefield, Massachusetts), the community standard SBML Level 3 Version 1 FBC Package Version 2 through libSBML [37] and YAML for easy tracking of differences between model files. Meanwhile, backwards compatibility ensures that Excel and SBML files generated by earlier RAVEN versions can still be imported.

From *de novo* draft GEMs to *Sco4*

An improved GEM of *S. coelicolor*, called *Sco4* for the fourth major published model, was generated through RAVEN 2.0 following the pipeline illustrated in Fig 3. The model is based on the complete genome sequences of *S. coelicolor* A3(2), including chromosome and two plasmids (GenBank accession: GCA_000203835.1) [23]. MetaCyc-based draft model was generated with `getMetaCycModelForOrganism` using default cut-offs (bit-score ≥ 100, positives ≥ 45%). Two KEGG-based draft models were generated with `getKEGGModelForOrganism` by (i) using 'sco' as KEGG organism identifier, and (ii) querying the *S. coelicolor* proteome against HMMs trained on prokaryotic sequences with 90% sequence identity. These two models were merged with `mergeModels`, subsequently combined with the MetaCyc-based draft using `combineMetaCycKEGGModels`, followed by manual curation. Reactions were mapped from iMK1208 to MetaCyc and KEGG identifiers in a semi-automated manner (S1 Table). Metabolites in iMK1208 were associated to MetaCyc and KEGG identifiers through examining the mapped reactions (S2 Table). Pathway gaps and invalid metabolite identifiers were thus detected and revised accordingly.

Manual curation of the combined draft and iMK1208 culminated in the *Sco4* model. Curation entailed identifying reactions from the combined draft, taking into account the absence of gene-associations in iMK1208; explicit subsystem and/or pathway information; support from both MetaCyc and KEGG reconstructions; additional literature information, as well as potential taxonomic conflicts. Manual curation was particularly required for secondary metabolite biosynthetic pathways, due to high levels of sequence similarity among the synthetic domains of polyketide synthase and nonribosomal peptide synthethase [38]. The identified new reactions were added to *Sco4*, while retaining the previous manual curation underlying iMK1208. Spontaneous reactions were added through `addSpontaneousRxns`, while transport reactions annotated in the MetaCyc-based reconstruction were manually curated. Gene essentiality was simulated on iMK1208 and *Sco4* by the COBRA function `singleGeneDeletion`, with a more than 75% reduction in growth rate aidentifying essential reactions. Potential targets for metabolic engineering were predicted using the flux scanning with enforced objective function `FSEOF` [30]. The reconstruction and curation of *Sco4* is provided as a MATLAB script in GitHub.

Model repository

The updated *Sco4* model is deposited to a GitHub repository in MATLAB .mat, SBML L3V1 FBCv2 .xml, Excel .xlsx, YAML .yml and flat-text .txt formats. Users can not only download the most recent version of the model, but also report issues and suggest changes. Updates in the metabolic network or gene associations can readily be tracked by querying the difference in the flat-text model

and YAML representations. As such, *Sco4* aims to be a community model, where improved knowledge and annotation will incrementally and constantly refine the model of *S. coelicolor*.

Acknowledgements

We thank Dr. Sylvain Prigent for valuable discussions.

Availability

RAVEN is an open source software package available in the GitHub repository https://github.com/SysBioChalmers/RAVEN. The updated *S. coelicolor* genome-scale metabolic model *Sco4* is available as a public GitHub repository at https://github.com/SysBioChalmers/Streptomyces_coelicolor-GEM. Supporting Information is available from: http://doi.org/10.6084/m9.figshare.6236903.

References

1. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 2010;5: 93–121. doi:10.1038/nprot.2009.203
2. O'Brien EJ, Monk JM, Palsson BO. Using genome-scale models to predict biological capabilities. Cell. Elsevier Inc.; 2015;161: 971–987. doi:10.1016/j.cell.2015.05.019
3. Simeonidis E, Price ND. Genome-scale modeling for metabolic engineering. J Ind Microbiol Biotechnol. 2015;42: 327–38. doi:10.1007/s10295-014-1576-3
4. 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. Nature Publishing Group; 2016;35: 81–89. doi:10.1038/nbt.3703
5. Mardinoglu A, Gatto F, Nielsen J. Genome-scale modeling of human metabolism - a systems biology approach. Biotechnol J. 2013;8: 985–996. doi:10.1002/biot.201200275
6. Nielsen J. Systems Biology of Metabolism: A Driver for Developing Personalized and Precision Medicine. Cell Metab. Elsevier Inc.; 2017;25: 572–579. doi:10.1016/j.cmet.2017.02.002
7. Hwang K-S, Kim HU, Charusanti P, Palsson BØ, Lee SY. Systems biology and biotechnology of Streptomyces species for the production of secondary metabolites. Biotechnol Adv. Elsevier Inc.; 2014;32: 255–268. doi:10.1016/j.biotechadv.2013.10.008
8. Kim WJ, Kim HU, Lee SY. Current state and applications of microbial genome-scale metabolic models. Curr Opin Syst Biol. Elsevier Ltd; 2017;2: 10–18. doi:10.1016/j.coisb.2017.03.001
9. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J. The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum. PLoS Comput Biol. 2013;9: e1002980. doi:10.1371/journal.pcbi.1002980
10. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45: D353–D361. doi:10.1093/nar/gkw1092
11. Thompson RA, Dahal S, Garcia S, Nookaew I, Trinh CT. Exploring complex cellular phenotypes and

model-guided strain design with a novel genome-scale metabolic model of Clostridium thermocellum DSM 1313 implementing an adjustable cellulosome. Biotechnol Biofuels. BioMed Central; 2016;9: 194. doi:10.1186/s13068-016-0607-x

12. Hamilton JJ, Calixto Contreras M, Reed JL. Thermodynamics and H<inf>2</inf> Transfer in a Methanogenic, Syntrophic Community. PLoS Comput Biol. 2015;11: 1–20. doi:10.1371/journal.pcbi.1004364
13. Shoaie S, Ghaffari P, Kovatcheva-Datchary P, Mardinoglu A, Sen P, Pujos-Guillot E, et al. Quantifying Diet-Induced Metabolic Changes of the Human Gut Microbiome. Cell Metab. 2015;22: 320–331. doi:10.1016/j.cmet.2015.07.001
14. Levering J, Broddrick J, Dupont CL, Peers G, Beeri K, Mayers J, et al. Genome-scale model reveals metabolic basis of biomass partitioning in a model diatom. PLoS One. 2016;11: 1–22. doi:10.1371/journal.pone.0155038
15. Chiappino-Pepe A, Tymoshenko S, Ataman M, Soldati-Favre D, Hatzimanikatis V. Bioenergetics-based modeling of Plasmodium falciparum metabolism reveals its essential genes, nutritional requirements, and thermodynamic bottlenecks. PLoS Comput Biol. 2017;13: 1–24. doi:10.1371/journal.pcbi.1005397
16. Sharma M, Shaikh N, Yadav S, Singh S, Garg P, Nowicki C, et al. A systematic reconstruction and constraint-based analysis of Leishmania donovani metabolic network: identification of potential antileishmanial drug targets. Mol BioSyst. Royal Society of Chemistry; 2017;277: 38245–38253. doi:10.1039/C6MB00823B
17. Ledesma-Amaro R, Kerkhoven EJ, Revuelta JL, Nielsen J. Genome scale metabolic modeling of the riboflavin overproducer Ashbya gossypii. Biotechnol Bioeng. 2014;111: 1191–1199. doi:10.1002/bit.25167
18. Agren R, Mardinoglu A, Asplund A, Kampf C, Uhlen M, Nielsen J. Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Mol Syst Biol. 2014;10: 1–13. doi:10.1002/msb.145122
19. Hur W, Ryu JY, Kim HU, Hong SW, Lee EB, Lee SY, et al. Systems approach to characterize the metabolism of liver cancer stem cells expressing CD133. Sci Rep. Nature Publishing Group; 2017;7: 45557. doi:10.1038/srep45557
20. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc. 2011;6: 1290–1307. doi:10.1038/nprot.2011.308
21. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat Protoc. 2007;2: 727–38. doi:10.1038/nprot.2007.99
22. 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–D480. doi:10.1093/nar/gkv1164
23. Bentley S, Chater K, Cerdeño-Tárraga A-M, Challis GL, Thomson NR, James KD, et al. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). Nature. 2002;417: 141–147. doi:10.1038/417141a

24. Wang H, Fewer DP, Holm L, Rouhiainen L, Sivonen K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. Proc Natl Acad Sci U S A. 2014;111: 9259–64. doi:10.1073/pnas.1401734111
25. Challis GL. Exploitation of the Streptomyces coelicolor A3(2) genome sequence for discovery of new natural products and biosynthetic pathways. J Ind Microbiol Biotechnol. 2014;41: 219–32. doi:10.1007/s10295-013-1383-2
26. Borodina I, Krabben P, Nielsen J. Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism. Genome Res. 2005;15: 820–9. doi:10.1101/gr.3364705
27. Alam MT, Merlo ME, Hodgson DA, Wellington EMH, Takano E, Breitling R. Metabolic modeling and analysis of the metabolic switch in Streptomyces coelicolor. BMC Genomics. 2010;11: 202. doi:10.1186/1471-2164-11-202
28. Kim M, Sang Yi J, Kim J, Kim J-N, Kim MW, Kim B-G. Reconstruction of a high-quality metabolic model enables the identification of gene overexpression targets for enhanced antibiotic production in Streptomyces coelicolor A3(2). Biotechnol J. 2014;9: 1185–94. doi:10.1002/biot.201300539
29. Bordel S, Agren R, Nielsen J. Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. PLoS Comput Biol. 2010;6: e1000859. doi:10.1371/journal.pcbi.1000859
30. Choi HS, Lee SY, Kim TY, Woo HM. In Silico Identification of Gene Amplification Targets for Improvement of Lycopene Production. Appl Environ Microbiol. 2010;76: 3097–3105. doi:10.1128/AEM.00115-10
31. Xu Z, Wang Y, Chater KF, Ou H, Xu HH, Deng Z, et al. Large-Scale Transposition Mutagenesis of Streptomyces coelicolor Identifies Hundreds of Genes Influencing Antibiotic Biosynthesis. Drake HL, editor. Appl Environ Microbiol. 2017;83: e02889-16. doi:10.1128/AEM.02889-16
32. Pabinger S, Rader R, Agren R, Nielsen J, Trajanoski Z. MEMOSys: Bioinformatics platform for genome-scale metabolic models. BMC Syst Biol. 2011;5: 20. doi:10.1186/1752-0509-5-20
33. Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, Nielsen J. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. PLoS Comput Biol. 2012;8. doi:10.1371/journal.pcbi.1002518
34. Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22: 1658–1659. doi:10.1093/bioinformatics/btl158
35. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780. doi:10.1093/molbev/mst010
36. Finn RD, Clements J, Eddy SR. HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res. 2011;39: 29–37. doi:10.1093/nar/gkr367
37. Bornstein BJ, Keating SM, Jouraku A, Hucka M. LibSBML: An API library for SBML. Bioinformatics. 2008;24: 880–881. doi:10.1093/bioinformatics/btn051
38. Medema MH, Blin K, Cimermancic P, De Jager V, Zakrzewski P, Fischbach MA, et al. AntiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 2011;39: 339–346. doi:10.1093/nar/gkr466

Tables

Table 1. Feature comparison of GEM reconstruction toolboxes.

		MEMO Sys	FAME	Microbes Flux	CoReCo	Pathway Tools	RAVEN 1.0	COBRA 3.0[a]	Model SEED	merlin	RAVEN 2.0
Reconstruct GEM for	Prokaryote	✓	✓	✓	✓	✓	✓		✓	✓	✓
	Eukaryote				✓	✓	✓			✓	✓
	Tissues/cell type						✓				✓
Reconstruct GEM based on	KEGG	✓	✓	✓	✓		✓		✓	✓	✓
	MetaCyc					✓					✓
	HMR										✓
	Template model[b]						✓		✓		✓
Feature	Mass and charge balance[c]				✓	✓		✓	✓	✓	✓
	Sub-cellular localization[d]					✓	✓	✓	✓	✓	✓
	Transporters annotation					✓			✓	✓	✓
	Spontaneous reaction					✓					✓
Import/ Export	SBML	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	YAML										✓
	Excel	✓	✓				✓	✓	✓		✓
	Model simulation		✓			✓	✓	✓			✓
	Pathways visualisation[e]		✓	✓		✓	✓	✓	✓	✓	✓

a. While COBRA does not support GEM reconstruction, it is included here as widely used MATLAB toolbox for GEM analysis.
b. An existing GEM can be used as temptlate for model reconstruciton based the sequence homology.
c. The mass- and charge-balanced reactions are derived from the MetaCyc database.
d. The sub-cellular localization of reactions can be estimated using the predictLocalization function in RAVEN.
e. RAVEN allows the visualize of simulation results by by overlaying information on pre-drawn metabolic maps.

Table 2. Identification of four metabolites with invalid KEGG identifiers.

Identifiers in iMK1208			Manual curation		Updated
Metabolite	KEGG	Reaction	MetaCyc reaction id	KEGG reaction id	KEGG
dtdp4d6dg	C00687	TDPDRE	DTDPDEHYDRHAMEPIM-RXN	R06514	C11907
		TDPGDH	DTDPGLUCDEHYDRAT-RXN	R06513	
r1p	C00442	PPM	PPENTOMUT-RXN	R01057	C00620
		PUNP1	ADENPHOSPHOR-RXN	R01561	
		PUNP3	RXN0-5199	R02147	
		PUNP5	INOPHOSPHOR-RXN	R01863	
		PUNP7	XANTHOSINEPHOSPHORY-RXN	R02297	
f1p	C02976	FRUK	1PFRUCTPHOSN-RXN	R02071	C01094
		FBA2	RXN-8631	R02568	
		FRUpts	R03232	R03232	
2hymeph	C06191	6PHBG	RXN0-5297	R05134	C02323
		DM_2HYMEPH	Assumed exchange reaction		

Figure legends

Fig 1. Schematic overview of RAVEN Toolbox version 2.0. RAVEN 2.0 was significantly enhanced for *de novo* draft model reconstruction by integrating the knowledge from different sources (e.g. KEGG and MetaCyc). It has now better import/export support to relevant formats, and especially improved compatibility with the COBRA toolbox. As a MATLAB toolbox, RAVEN 2.0 provides one unified environment both for model reconstruction and simulation analysis (e.g. FSEOF), and allows scripting for more flexible operations.

Fig 2. Schematic illustration of *Sco4* model generation by RAVEN 2.0. Three draft GEMs were automatically reconstructed in *de novo* based on MetaCyc and KEGG databases. The two GEMs derived from KEGG genome annotation and HMMs were merged and then subsequently combined with the MetaCyc-based model. The reactions and annotations in the combined draft model were used for manual curation of iMK1208 to result in *Sco4*.

Fig 3. Venn diagrams comparing genes and metabolites between the three *de novo* draft reconstructions and iMK1208.

Fig 4. The compositional subsystem distribution of the new reactions selected from the combined draft model. (A) A total of 398 metabolic reactions were introduced in *Sco4*, thereby expanding the coverage in both primary and secondary metabolism. (B) These 398 curated reactions were identified through either KEGG- and/or MetaCyc-based *de novo* reconstructions. (C) Metabolites from these new reactions are closely associated with the previous *S. coelicolor* GEM. Over 95% of the reactions have one or more metabolites included in iMK1208, and 36 reactions have all reactants and products included in the previous model, resulting in increased connectivity of the network; while only 19 reactions have all newly introduced metabolites, representing completely new stoichiometric network reconstructed by RAVEN 2.0.

Fig 5. Growth and gene essentiality prediction of *Sco4*. (A) Comparison of experimental growth rate and predicted growth rate of published bioreactor cultivations [31], using reported exchange rates as model constraints. (B) Comparison between the previously curated GEM iMK1208 and *Sco4*, of network characteristics and results from single-gene non-essentiality predictions, based on a mutagenesis study by [32].

Fig 6. Venn diagrams of targets predicted to be beneficial for overproduction of various secondary metabolites by FSEOF. The 7 secondary metabolites that *Sco4* can produce are grouped (A-C), and the overlap of target reactions are indicated within each group. Reactions that overlap within each group are subsequently compared between all groups (D), including in which pathways these reactions are involved. BCAA: branch-chain amino acid; TPI: triosephosphate isomerase; MEP/DOXP: mevalonate independent pathway.

Fig 7. Flow-chart of the functions for MetaCyc-based *de novo* GEM reconstruction in RAVEN 2.0. The MetaCyc-based module (indicated by the dashed line) reads in data files of MetaCyc pathways, reactions, compounds and enzymes, and converts the information into a model structure, which can then be utilized for automatic draft GEM generation through querying the protein sequences in MetaCyc (protseq.fsa). Different components are color coded as MetaCyc data files in blue, MATLAB functions in red, and data structures in green.

Supporting information

Available from: http://doi.org/10.6084/m9.figshare.6236903

S1 Data. Model structure after merging three *de novo* draft reconstructions.

S1 Table. Mapping of reactions to *de novo* reconstructions. To facilitate comparison of the existing iMK1208 model and the *de novo* reconstructions, iMK1208 reactions are mapped to their respective KEGG and MetaCyc counterparts.

S2 Table. Association of metabolites to KEGG and MetaCyc identifiers. To facilitate comparison of the existing iMK1208 model and the *de novo* reconstructions, iMK1208 metabolites are matched to their respective KEGG and MetaCyc identifiers.

S3 Table. Reactions from the *de novo* draft models used in *Sco4* generation. A set of reactions was selected from the *de novo* draft model for manual curation of the existing model towards *Sco4*.

S4 Table. Missing gene-associations in that were filled by *de novo* reconstruction. These reactions had no gene associated in iMK1208, but *de novo* reconstruction with RAVEN 2.0 identified the responsible genes.

S5 Table. Spontaneous reactions included in *Sco4* model, as identified from MetaCyc.

S6 Table. Transport reactions added to *Sco4* model based on MetaCyc-based reconstruction and manual curation.

S7 Table: Complete list of invalid KEGG identifiers and their replacements.

S8 Table. Comparison of growth media validation results between iMK1208 and Sco4. The growth media validation on iMK1208 and Sco4 models. The culture and simulating conditions were described previously [28]. Predictions on iMK1208 and Sco4 showed identical results.

S9 Table: Identification of putative essential genes from mutagenesis study and predictions from the iMK1208 and *Sco4* models. Logical indicating whether a gene has been identified in any of the three analyses.

S10 Table. Combined FSEOF results on the secondary metabolites included in *Sco4* model. Target reactions identified for each of the secondary metabolites are indicated, together their gene association. Direction indicates whether a reaction is reversible and running in the forward direction (1), running in the reverse direction (-1), or is irreversible (0).

S11 Table. Definition of RAVEN model structure fields. Where applicable, corresponding COBRA model structure fields are indicated.

S12 Table. Reaction identifier correspondence between KEGG and MetaCyc. The mapping information was provided by MetaCyc and expanded using function `linkMetaCycKEGGRxns`.

S13 Table. Metabolite identifier correspondence between KEGG and MetaCyc. The mapping information was provided by MetaCyc.

Figure 1

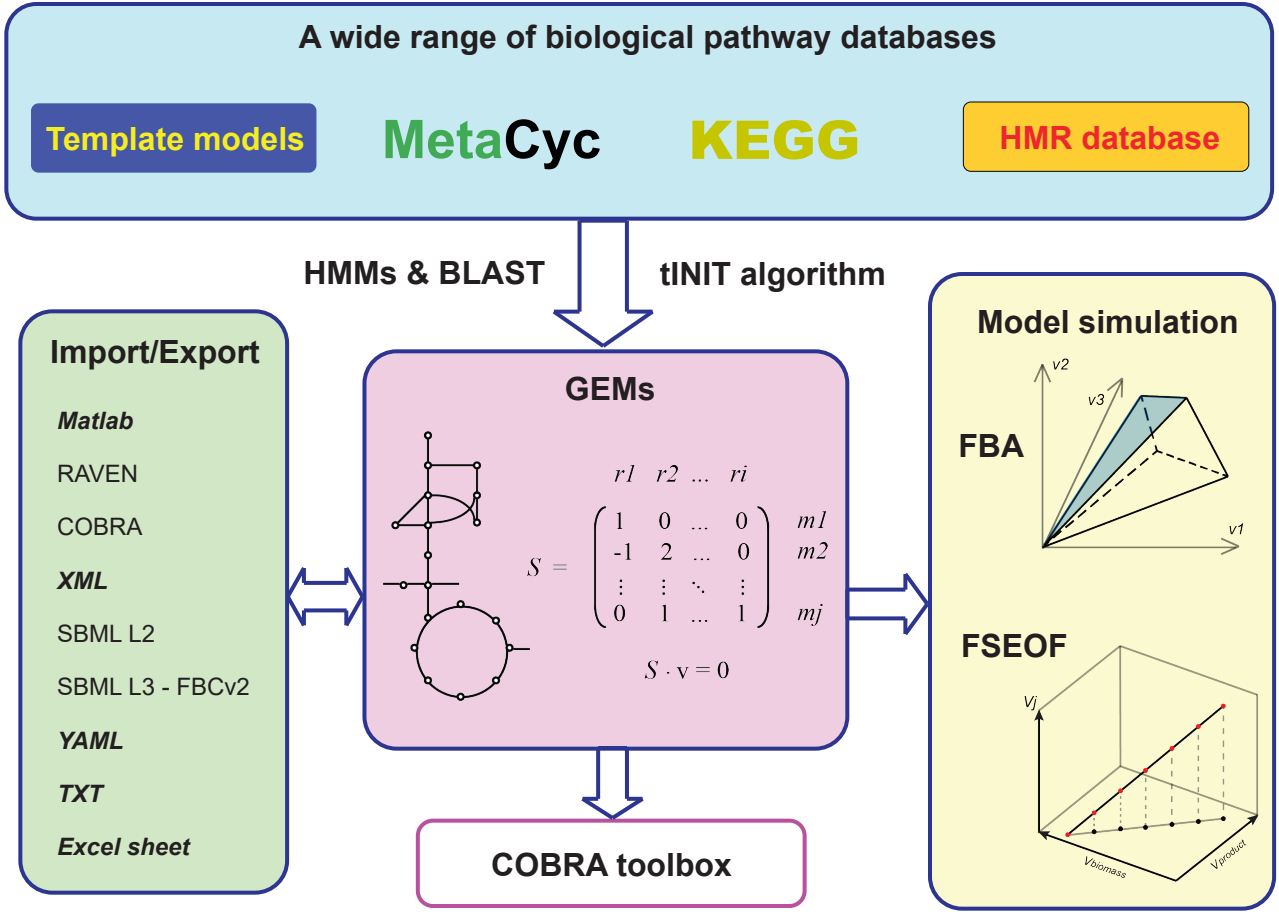

A wide range of biological pathway databases
Template models
MetaCyc
KEGG
HMR database
HMMs & BLAST
tINIT algorithm
Import/Export
Matlab
RAVEN
COBRA
XML
SBML L2
SBML L3 - FBCv2
YAML
TXT
Excel sheet
GEMs
r1 r2 ... ri
S =
1 0 ... 0
-1 2 ... 0
0 1 ... 1
m1
m2
mj
S · v = 0
COBRA toolbox
Model simulation
FBA
v2
v3
v1
FSEOF
Vj
Vbiomass
Vproduct

Figure 2

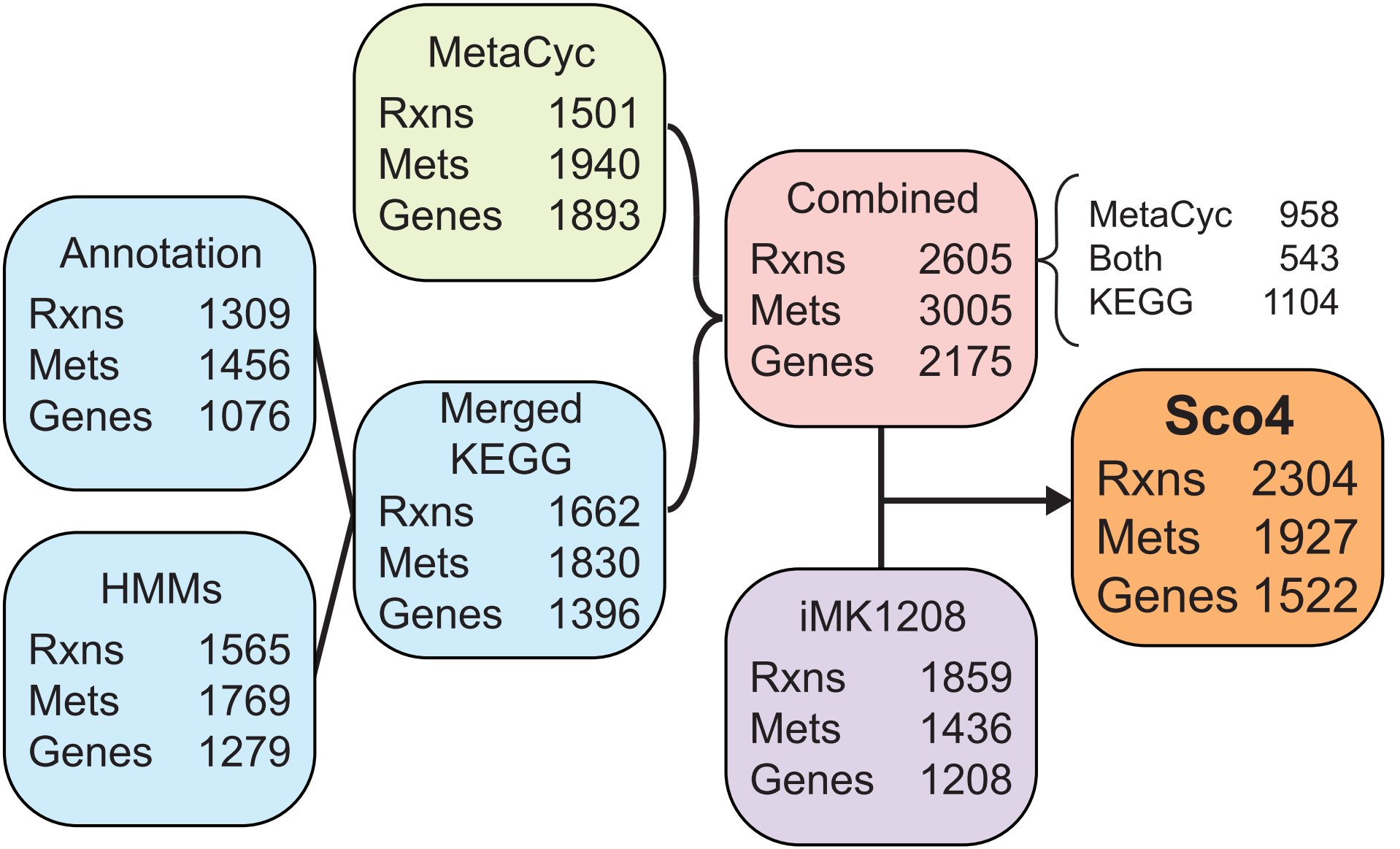

MetaCyc
Rxns 1501
Mets 1940
Genes 1893
Annotation
Rxns 1309
Mets 1456
Genes 1076
HMMs
Rxns 1565
Mets 1769
Genes 1279
Merged KEGG
Rxns 1662
Mets 1830
Genes 1396
Combined
Rxns 2605
Mets 3005
Genes 2175
MetaCyc 958
Both 543
KEGG 1104
iMK1208
Rxns 1859
Mets 1436
Genes 1208
Sco4
Rxns 2304
Mets 1927
Genes 1522

Figure 3

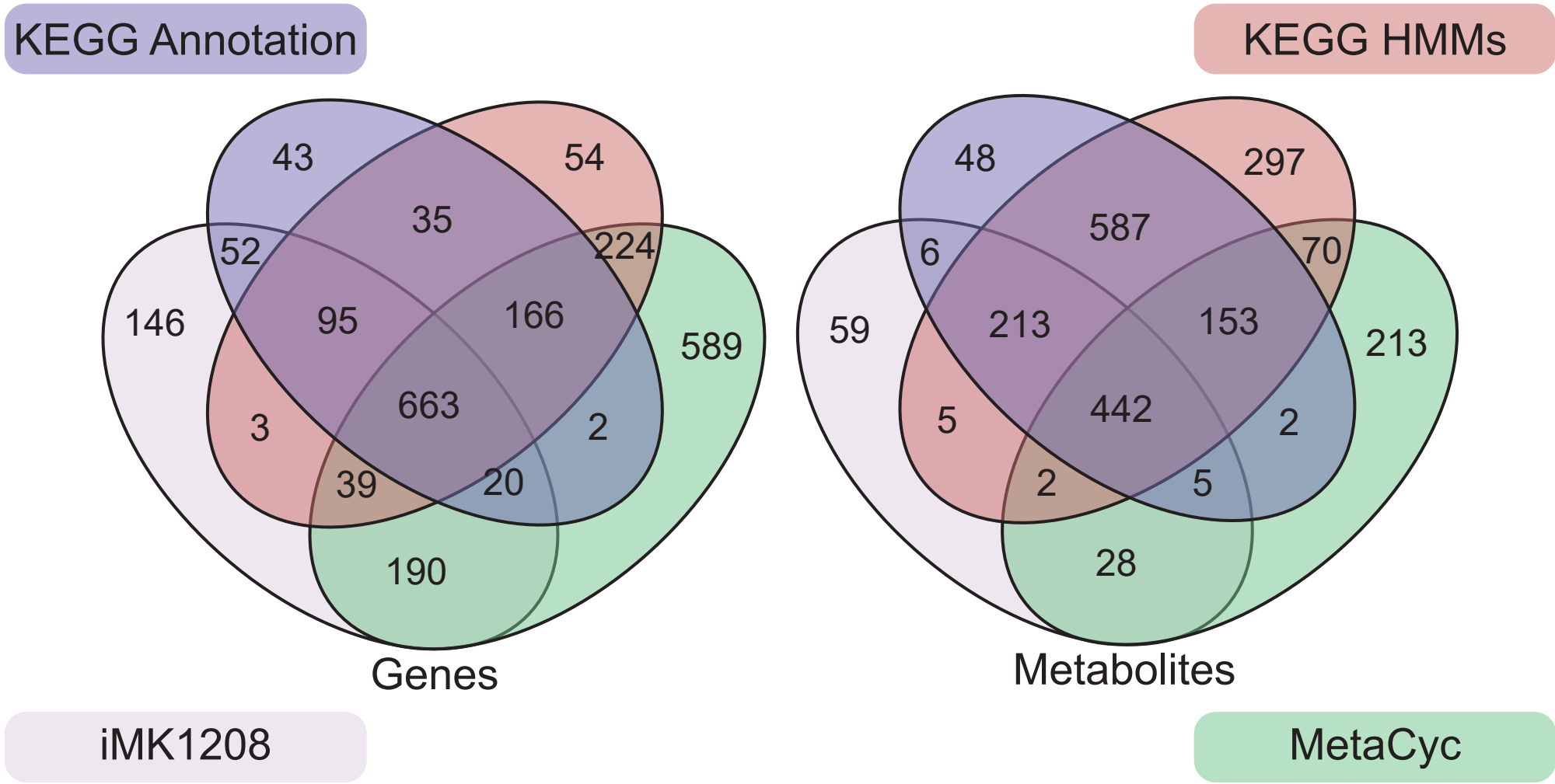

KEGG Annotation
KEGG HMMs
43
54
35
52
224
146
95
166
589
663
3
2
39
20
190
Genes
48
297
587
6
70
59
213
153
213
442
5
2
2
5
28
Metabolites
iMK1208
MetaCyc

Figure 4

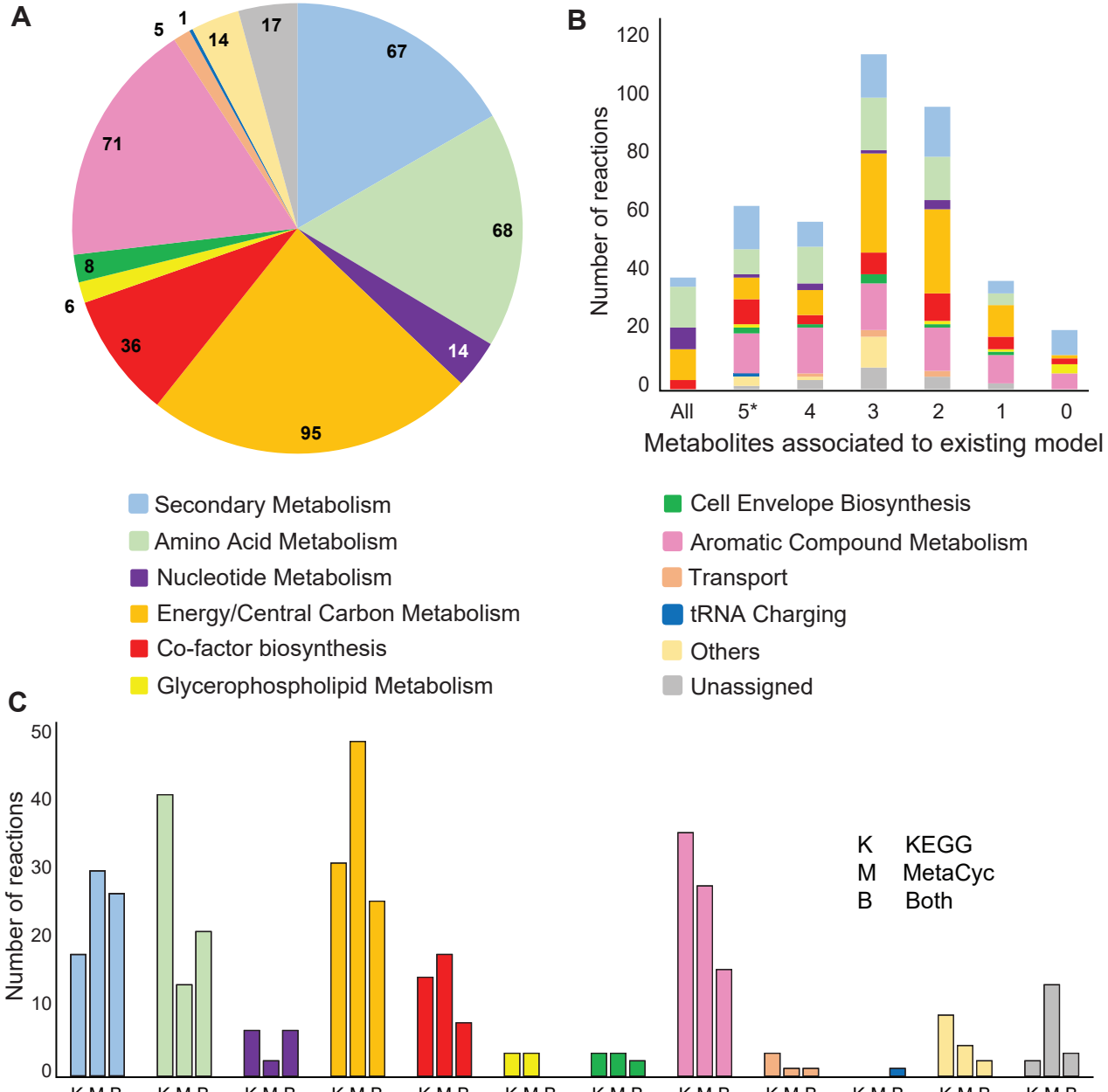
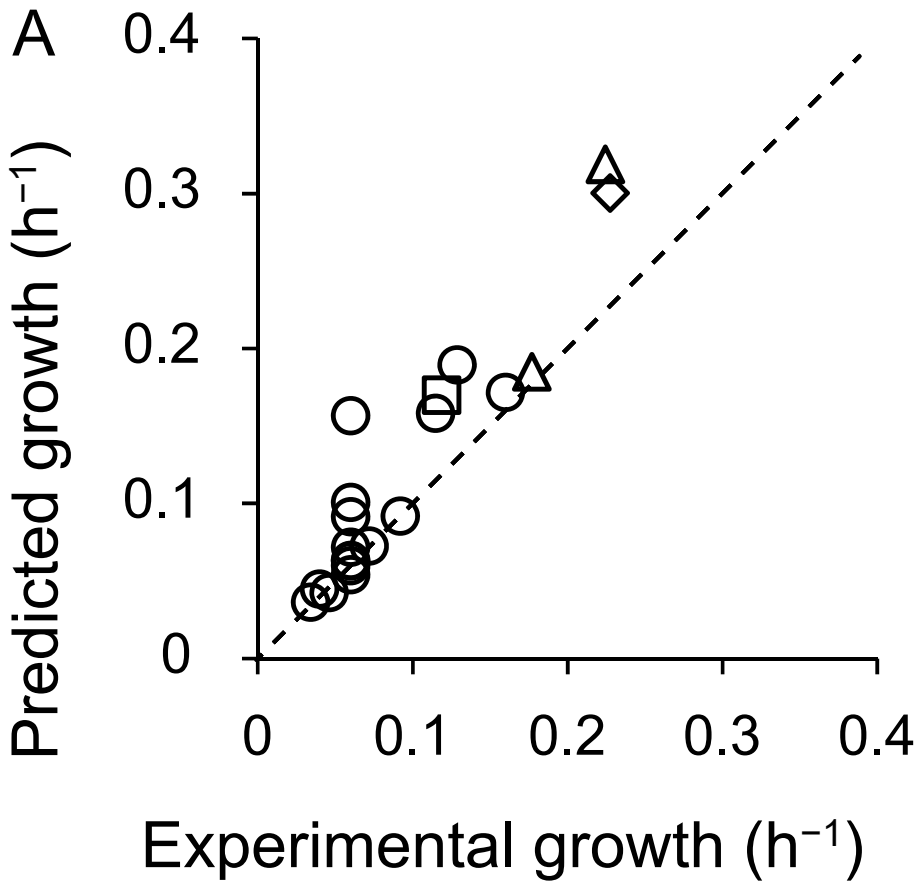

A
67
68
14
95
36
6
8
71
5
1
14
17
B
Number of reactions
120
100
80
60
40
20
0
All
5*
4
3
2
1
0
Metabolites associated to existing model
Secondary Metabolism
Amino Acid Metabolism
Nucleotide Metabolism
Energy/Central Carbon Metabolism
Co-factor biosynthesis
Glycerophospholipid Metabolism
Cell Envelope Biosynthesis
Aromatic Compound Metabolism
Transport
tRNA Charging
Others
Unassigned
C
Number of reactions
50
40
30
20
10
0
K M B
K KEGG
M MetaCyc
B Both

Figure 5

A



△ Nieselt *et al.* 2010
◇ Waldvogel *et al.* 2011
○ Melzoch *et al.* 1997
□ Thomas *et al.* 2012

B

		iMK1208	Sco4
	Reactions	1859	2299
	Metabolites	1436	1924
	Genes	1209	1511
Essentiality prediction	Sensitivity	0.508	0.462
	Specificity	0.876	0.901
	Accuracy	0.858	0.882

Figure 6

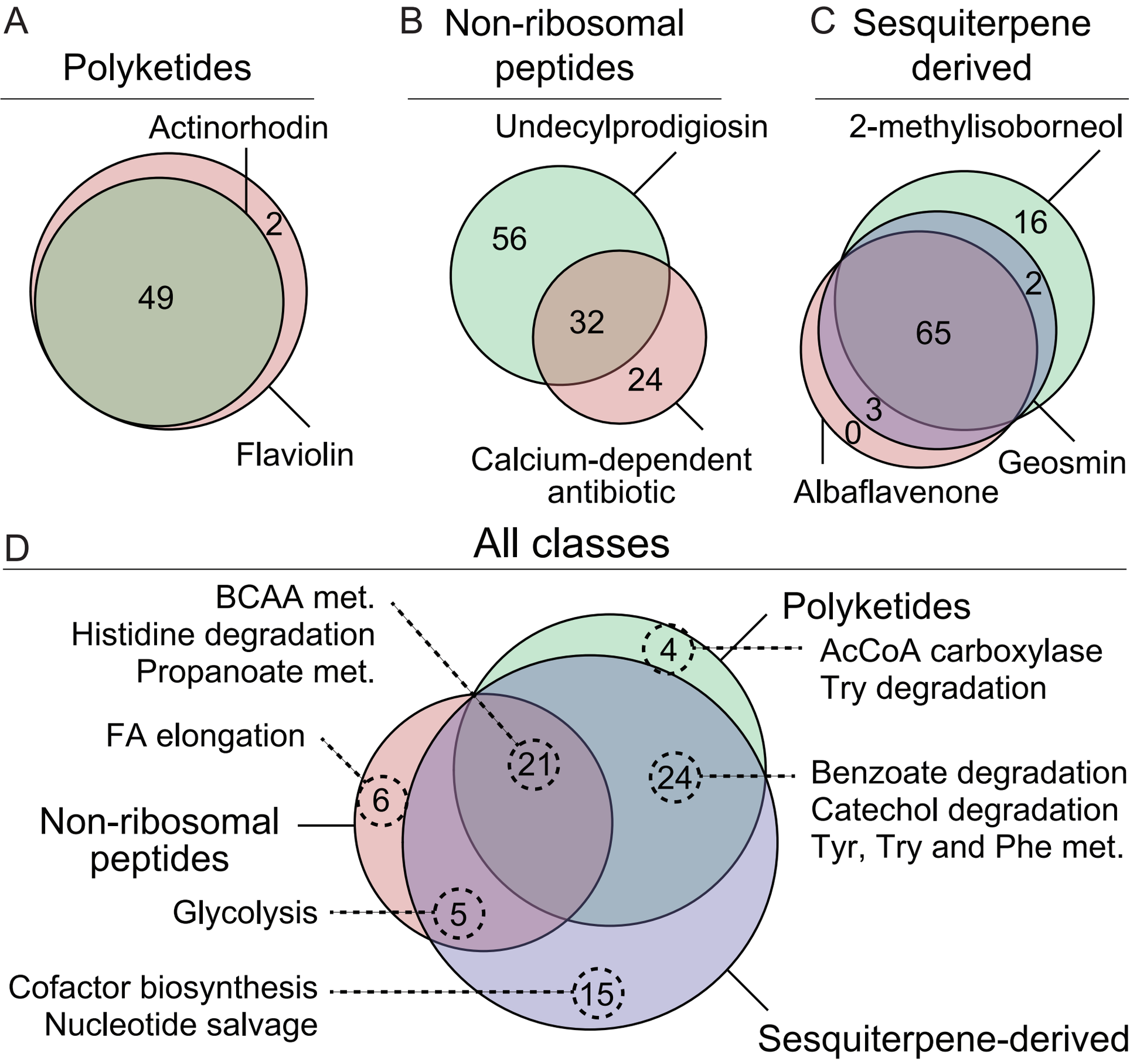
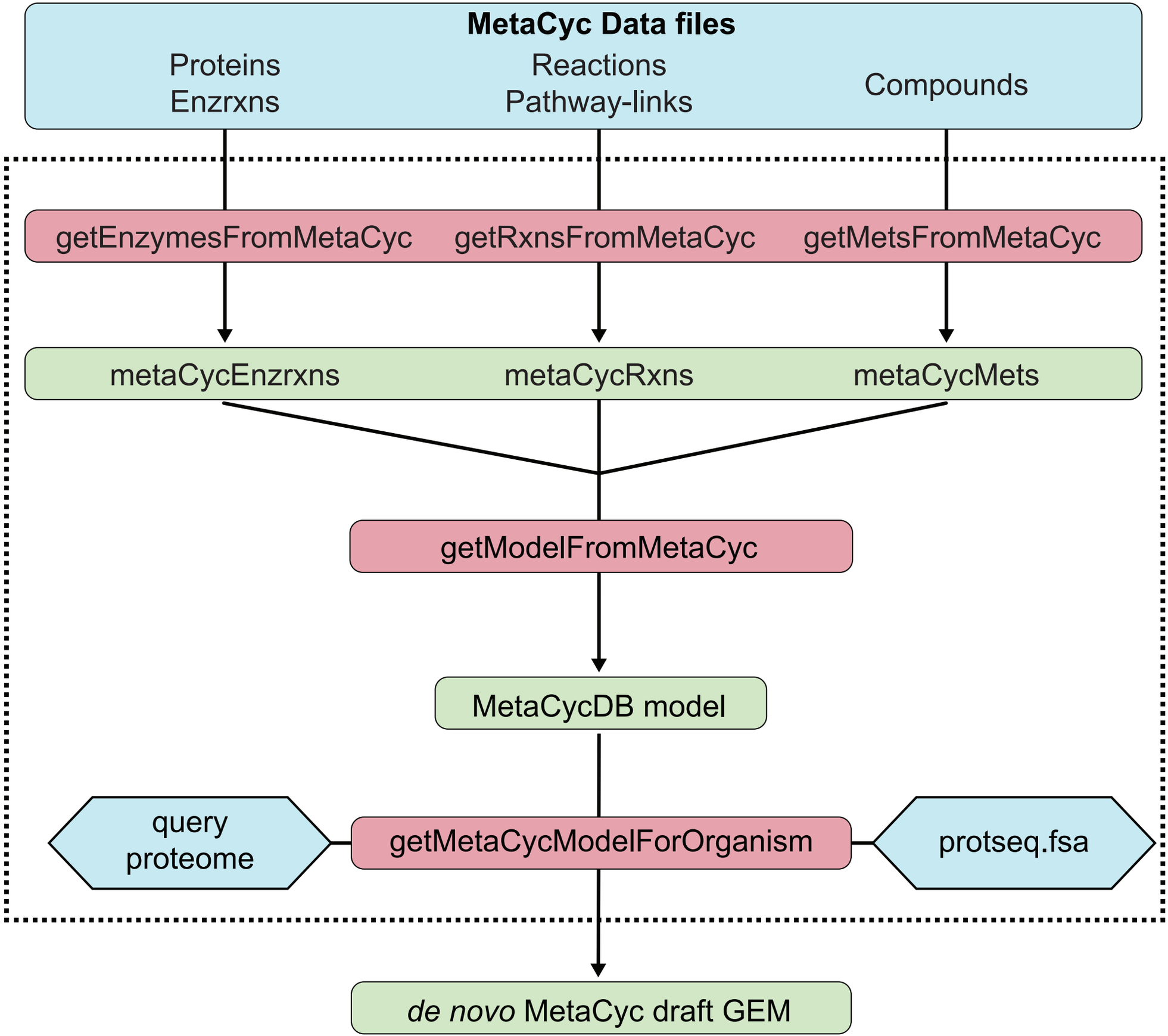

A
Polyketides
Actinorhodin
2
49
Flaviolin
B
Non-ribosomal peptides
Undecylprodigiosin
56
32
24
Calcium-dependent antibiotic
C
Sesquiterpene derived
2-methylisoborneol
16
2
65
3
0
Geosmin
Albaflavenone
D
All classes
BCAA met.
Histidine degradation
Propanoate met.
Polyketides
4
AcCoA carboxylase
Try degradation
FA elongation
21
24
Benzoate degradation
Catechol degradation
Tyr, Try and Phe met.
6
Non-ribosomal peptides
Glycolysis
5
Cofactor biosynthesis
Nucleotide salvage
15
Sesquiterpene-derived

Figure 7


MetaCyc Data files
Proteins
Enzrxns
Reactions
Pathway-links
Compounds
getEnzymesFromMetaCyc
getRxnsFromMetaCyc
getMetsFromMetaCyc
metaCycEnzrxns
metaCycRxns
metaCycMets
getModelFromMetaCyc
MetaCycDB model
query proteome
getMetaCycModelForOrganism
protseq.fsa
de novo MetaCyc draft GEM