Title

MAGI: A method for metabolite, annotation, and gene integration

Authors

Erbilgin, Onur

Rübel, Oliver

Louie, Katherine B

et al.

Publication Date

2017

DOI

10.1101/204362

Peer reviewed

# MAGI: A method for metabolite, annotation, and gene integration

Onur Erbilgin[1], Oliver Rübel[2], Katherine B. Louie[3], Matthew Trinh[1], Markus de Raad[1], Tony Wildish[3,4], Daniel Udwary[3,4], Cindi Hoover[3], Samuel Deutsch[1,3], Trent R. Northen[1,3,*], Benjamin P. Bowen[1,3,*]

1. Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory
2. Data Analytics and Visualization Group, Computational Research Division, Lawrence Berkeley National Laboratory
3. Joint Genome Institute, Lawrence Berkeley National Laboratory
4. National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory

## Author Contributions

OE, BPB, TRN conceived and designed the method
OE, BPB, OR, MT, TW, DWU developed the method
KBL, MdR, CAH conducted the experiments
OE, BPB, SD, TRN, KBL wrote the manuscript.

***Correspondence** should be addressed to bpbowen@lbl.gov and trnorthen@lbl.gov

## Abstract

Metabolomics is a widely used technology for obtaining direct measures of metabolic activities from diverse biological systems. However, ambiguous metabolite identifications are a common challenge and biochemical interpretation is often limited by incomplete and inaccurate genome-based predictions of enzyme activities (*i.e.* gene annotations). Metabolite, Annotation, and Gene Integration (MAGI) generates a metabolite-gene association score using a biochemical reaction network. This is calculated by a method that emphasizes consensus between metabolites and genes via biochemical reactions. To demonstrate the potential of this method, we applied MAGI to integrate sequence data and metabolomics data collected from *Streptomyces coelicolor* A3(2), an extensively characterized bacterium that produces diverse secondary metabolites. Our findings suggest that coupling metabolomics and genomics data by scoring consensus between the two increases the quality of both metabolite identifications and gene annotations in this organism. MAGI also made biochemical predictions for poorly annotated genes that were consistent with the extensive literature on this important organism. This limited analysis suggests that using metabolomics data has the potential to improve annotations in sequenced organisms and also provides testable hypotheses for specific biochemical functions. MAGI is freely available for academic use both as an online tool at https://magi.nersc.gov and with source code available at https://github.com/biorack/magi.

## Introduction

Metabolomics approaches now enable global profiling, comparison, and discovery of diverse metabolites present in complex biological samples[1]. Liquid chromatography coupled with electrospray ionization mass spectrometry (LC-MS) is one of the leading methods in metabolomics[1]. A critical measure in metabolomics datasets is known as a “feature,” which is a unique combination of mass-to-charge (*m/z*) and chromatographic retention time[1]. Each distinct feature may match to hundreds of unique chemical structures. This makes metabolite identification (the accurate assignment of the correct chemical structure to each feature) one of the fundamental challenges in metabolomics[2-4]. To aid metabolite identification efforts, ions (each with a unique *m/z* and retention time) are typically fragmented, and the resulting fragments are compared against either experimental[5, 6] or computationally predicted[5, 7-11] reference libraries. While this method is highly effective at reducing the search space for metabolite identification, misidentifications are inevitable, especially for metabolites lacking authentic standards.

One strategy for addressing the large search space of compound identifications is to assess identifications in the context of the predicted metabolism of the organism(s) being studied. Several tools do this with varying degrees of complexity with strategies including directly mapping metabolites onto reactions[12] or scoring the likelihood of metabolite identities using reaction networks and predictive pathway mapping[13]. However, many metabolites cannot be included in these approaches. This is due to a number of factors, including the low coverage in reaction databases[14, 15] (especially for secondary metabolites[16-19]), incomplete or inaccurate set of reactions for an organism, and enzyme promiscuity not being taken into account when formulating the potential metabolism of an organism. To help address these challenges computational strategies have been developed including MyCompoundID [20, 21], IIMDB [22], MINES [23] and the ATLAS of biochemistry [24] to enzymatically enlarge compound space similar to the retrosynthesis tools such as Retrorules [25] and rePrime[26]. These approaches can be complimented by chemical networking to help address the limited number of metabolites represented in reactions, by expanding reaction space based on chemical or spectral similarity between metabolites. Effectively, even when a metabolite is not directly involved in a reaction, a linkage can still be made with a reaction based on similarity to another well-studied metabolite[16-19, 27]. In this way, chemical networking is a viable solution that expands reaction databases to integrate with already expansive metabolite databases. This allows more putative metabolite identifications to be assessed using the predicted metabolism of the organism(s).

Recently, approaches have been developed that span the gap between metabolomics and genomics and allow for some enzyme promiscuity. GNP, developed specifically for discovering new nonribosomal peptides (NRPs) and polyketides, uses a gene-forward strategy that predicts possible chemical structures of NRP and polyketide synthases and generates a set of predicted MS/MS spectra based on those predictions; these predictions are then used to mine MS data [28]. Pep2Path, also developed exclusively for NRPs and post-translationally modified peptides (RiPPs), takes a Bayesian approach to scoring putative NRPs and RiPPs based on the gene sequences present in the assayed organism [29]. Finally, a more general approach has been developed where a mutant library of an organism is assayed for major differences in the mass spectrometry profile, and the major differences are manually annotated with human intuition [30].

Due to the vast amount of knowledge about Streptomyces species, they are an excellent target for developing new tools for metabolite and genome exploration. Representatives from this genus produce many antibacterials, anticancer compounds, immunosuppresents, antifungals, cardiovascular agents, and veterinary products including erythromycin, tetracycline, doxorubicin, enediyenes, FK-506, rapamycin, avermectin, nemadectin, amphotericin, griseofulvin, nystatin, lovastatin, compactin, monensin, and tylosin [31]. Thus making them a highly relevant group for in depth studies to link natural products with associated genes. In particular, *Streptomyces coelicolor* is a model actinomycete secondary metabolite producer [32]; studies from over three thousand papers and over 60 years of work [33] have produced, among other things, a detailed understanding of the secondary metabolites this organism produces, where two are the pigmented antibiotics: actinorhodin and undecylprodiosin. These experiments have identified the biochemical pathways, genes, and regulatory processes that are necessary for producing the associated secondary metabolites [34].

Here we report Metabolite, Annotation, and Gene Integration (MAGI), an approach to generate metabolite-gene associations (Figure 1) by scoring consensus between metabolite identifications and gene annotations. MAGI is guided by the principles that the probability of a metabolite identity increases if there is genetic evidence to support that metabolite and that the probability of a gene function increases if there is metabolomic evidence for that function. Inputs to MAGI are typically a metabolite identification file of LCMS features and a protein or gene sequence FASTA file. For each LCMS feature, there are often many plausible metabolite identifications that can be given a probability based on accurate mass error and/or mass fragmentation comparisons. MAGI links these putative compound identifications to reactions both directly and indirectly by a biochemically relevant chemical similarity network. Likewise, MAGI associates input sequences to biochemical reactions by assessing sequence homology to reference sequences in the MAGI reaction database. For each sequence, there are often several plausible reactions with equal or similar probability. While annotation services would typically reduce specificity in these cases (*e.g., by* simply annotating as oxidoreductase), MAGI maintains all specific reactions as possibilities. Since MAGI links both metabolites and sequences to reactions with numerical scores that are proxies for probabilities, a final integrative MAGI score is calculated that magnifies consensus between a gene annotation and a metabolite identification. We applied this approach to one of the best characterized secondary metabolite producing bacteria, *Streptomyces coelicolor* A3(2)[35], by integrating its genome sequence with untargeted metabolomics data. MAGI successfully reduced the metabolite identity search space by scoring metabolite identities based on the predicted metabolism of an organism. Additionally, further investigation of the metabolite-gene associations led to identification of unannotated and misannotated genes that were subsequently validated using literature searches. This simple example illustrates the key aspects of MAGI.

## Methods

**Media and culture conditions.** A 20 µL volume of glycerol stock of wild-type *S. coelicolor* spores was cultured in 40 mL R5 medium in a 250-mL flask. One liter of R5 medium base included 103 g sucrose, 0.25 g $K_2SO_4$, 10.12 g $MgCl_2 \bullet 6H_20$, 10 g glucose, 0.1 g cas-amino acids, 2 mL trace element solution, 5 g yeast extract, and 5.73 g TES buffer to 1 L distilled water. After autoclave sterilization, 1 mL 0.5% $KH_2PO_4$, 0.4 mL 5M $CaCl_2 \bullet 2H_20$, 1.5 mL 20% L-proline, 0.7 ml 1N NaOH were added as per the following protocol: https://www.elabprotocols.com/protocols/#!protocol=486. Each flask contained a stainless steel spring (McMaster-Carr Supply, part 9663K77), cut to fit in a circle in the bottom of the flask. The spring was used to prevent clumping of *S. coelicolor* during incubation. A foam stopper was used to close each flask (Jaece Industries Inc., Fisher part 14-127-40D). Four replicates of each sample were grown in a 28°C incubator with shaking at 150 rpm. On day six, 1 mL from each replicate were collected in 2 mL Eppendorf tubes in a sterile hood. Samples were centrifuged at 3,200 x g for 8 minutes at 4 °C to pellet the cells. Supernatants were decanted into fresh 2 mL tubes and frozen at -80 °C. Pellets were flash frozen on dry ice and then stored at -80 °C.

**LCMS sample preparation and data acquisition.** In preparation for LCMS, medium samples were lyophilized. Dried medium was then extracted with 150 µL methanol containing an internal standard (2-Amino-3-bromo-5-methylbenzoic acid, 1 µg/mL, Sigma, #631531), vortexed, sonicated in a water bath for 10 minutes, centrifuged at 5,000 rpm for 5 min, and supernatant finally centrifuge-filtered through a 0.22 µm PVDF membrane (UFC40GV0S, Millipore). LC-MS/MS was performed in negative ion mode on a 2 µL injection, with UHPLC reverse phase chromatography performed using an Agilent 1290 LC stack and Agilent C18 column (ZORBAX Eclipse Plus C18, Rapid Resolution HD, 2.1 x 50 mm, 1.8 µm) at 60 °C and with MS and MS/MS data collected using a QExactive Orbitrap mass spectrometer (Thermo Scientific, San Jose, CA). Chromatography used a flow rate of 0.4 mL/min, first equilibrating the column with 100% buffer A (LC-MS water with 0.1% formic acid) for 1.5 min, then diluting over 7 minutes to 0% buffer A with buffer B (100% acetonitrile with 0.1% formic acid). Full MS spectra were collected at 70,000 resolution from *m/z* 80-1,200, and MS/MS fragmentation data collected at 17,500 resolution using an average of 10, 20 and 30 eV collision energies.

**Feature detection**. MZmine (version 2.23) [36] was used to deconvolute mass spectrometry features. The methods and parameters used were as follows (in the order that the methods were applied). MS/MS peaklist builder: retention time between 0.5-13.0 minutes, m/z window of 0.01, time window of 1.00. Peak extender: *m/z* tolerance 0.01 *m/z* or 50.0 ppm, min height of 1.0E0. Chromatogram deconvolution: local minimum search algorithm where chromatographic threshold was 1.0%, search minimum in RT range was 0.05 minutes, minimum relative height of 1.0%, minimum absolute height of 1.0E5, minimum ratio of peak top/edge of 1.2, peak duration between 0.01 and 30 minutes. Duplicate peak filter: m/z tolerance of 0.01 m/z or 50.0 ppm, RT tolerance of 0.15 minutes. Isotopic peaks grouper: *m/z* tolerance of 1.0E-6 m/z or 20.0 ppm, retention time tolerance of 0.01, maximum charge of 2, representative isotope was lowest *m/z*. Adduct search: RT tolerance of 0.01 minutes, searching for adducts M+Hac-H, M+Cl, with an *m/z* tolerance of 1.0E-5 *m/z* or 20.0 ppm and max relative adduct peak height of 1.0%. Join aligner: *m/z* tolerance of 1.0E-6 *m/z* or 50.0 ppm, weight for *m/z* of 5, retention time tolerance of 0.15 minutes, weight for RT of 3. Same RT and *m/z* range gap filler: *m/z* tolerance of 1.0E-6 *m/z* or 20.0 ppm.

**Metabolite identification.** During the LCMS acquisition, two MS/MS spectra were acquired for every MS spectrum. These MS/MS spectra are acquired using data-dependent criteria in which the 2 most intense ions are pursued for fragmentation, and then the next 2 most intense ions such that no ion is fragmented more frequently than every 10 seconds. To assign probable metabolite identities to a spectrum a modified version of the previously described MIDAS approach was used[7]. Our metabolite database is the merger of HMDB, MetaCyc, ChEBI, WikiData, GNPS, and LipidMaps resulting in approximately 180,000 unique chemical structures. For each of these structures, a comprehensive fragmentation tree was pre-calculated to a depth of 5 bond-breakages; these trees were used to accelerate the MIDAS scoring process. The source code to generate trees and score spectra against trees is available on GitHub (https://github.com/biorack/pactolus). The following procedure was used in the MIDAS scoring. Precursor *m/z* values were neutralized by 1.007276 Da. For each metabolite within 10 ppm of the neutralized precursor mass, MS/MS ions were associated with nodes of the fragmentation tree using a window of 0.01 Da using MS/MS neutralizations of 1.00727, 2.01510, and -0.00055, as described [7]. For metabolite-features of interest discussed in the text, retention time, m/z, adduct, and fragmentation pattern were used to define a Metabolite Atlas [37] library (Supplementary Data 1). For each metabolite, raw data was inspected manually using MZmine [36] to rule out peak misidentifications due to adduct formation and in-source degradation.

**MAGI biochemical reaction and reference sequence database.** The MAGI biochemical reaction database was constructed by aggregating all publicly available biochemical reactions in MetaCyc and RHEA biochemical reaction databases [14, 15]. This reaction database currently includes 12,293 unique metabolite structures. Identical reactions were collapsed together by calculating a “reaction InChI key,” where the SMILES strings of all members of a reaction were joined together, separated by a “.” and converted to a single InChI string through an RDkit (https://github.com/rdkit/rdkit) Mol object, and then the InChI key was calculated also using RDKit. Biochemical reactions with identical reaction InChI keys have identical chemical metabolites, indicating they are duplicates, and were collapsed into one database entry, retaining reference sequences. Reference sequences for each biochemical reaction from each database were combined to create a set of curated reference sequences for each biochemical reaction in the database.

**Chemical Network.** In order to expand the chemical space beyond what is in the biochemical reaction database, a chemical network was constructed to relate all metabolites in the database to metabolites in biochemical reactions by biochemical similarity. In each molecule, 70 chemical features (Supplementary Table 1) were located. These features were defined previously as being biochemically relevant [38]. The count of each feature was stored as a vector for each molecule. The Euclidean distance between two vectors was used to determine similarity between two molecules and construct a similarity network where every molecule is connected to every molecule by the difference in their vectors. This network was trimmed by calculating a minimum-spanning tree based on frequency of biochemical differences where more frequent differences would be preserved when possible (Supplementary Data 2).

**Gene Annotations of *Streptomyces coelicolor*.** KEGG annotations were obtained by submitting the S. coelicolor protein FASTA obtained from IMG to the KEGG Automatic Annotation Server version 2.1 [39] and downloading the gene-KO results table. KO numbers were associated with reactions by assessing if there was a link to one or more KEGG reaction entries directly from the webpage of that KO. For BioCyc annotations and reactions, the BioCyc *S. coelicolor* database was downloaded. For the reactions in Table 1, KEGG and BioCyc reactions were manually inspected and compared to MAGI reactions.

**MAGI workflow.** An input metabolite structure is expanded to similar metabolite structures as suggested by the chemical network and all tautomers of those metabolites. Searching all tautomeric forms of a metabolite structure is a known method to enhance metabolite database searches [40]. Tautomers were generated by using the MolVS package. The reaction database is then queried to find reactions containing these metabolites or their tautomers. Direct matches are stereospecific, but tautomer matches are not. This is due to limitations in the tautomer generating method and in how the chemical network was constructed. The metabolite score, *C*, is inherited from the MS/MS scoring algorithm and is a proxy for the probability that a metabolite structure is correctly assigned. In our case, it is the MIDAS score, but could be any score due to the use of the geometric mean to calculate the MAGI score. The metabolite score is set to 1 as a default.

If the reaction has a reference sequence associated with it, this reference sequence is used as a BLAST query against a sequence database of the input gene sequences to find genes that may encode that reaction. The reciprocal BLAST is also performed, where genes in the input gene sequences are queries against the reaction reference sequence database; this finds the reactions that a gene may encode for. The BLAST results are joined by their common gene sequence and are used to calculate a homology score: $H=F+R-|F-R|$ where *F* and *R* are log-transformed e-values of the BLAST results (a proxy for the probability that two gene sequences are homologs), with *F* representing the reaction-to-gene BLAST score, and *R* the gene-to-reaction BLAST score. The homology score is set to 1 if no sequence is matched.

The reciprocal agreement between both BLAST searches is also assessed, namely whether they both agreed on the same reaction or not, formulating a reciprocal agreement score: $\alpha$. $\alpha$ is equal to 2 for reciprocal agreements, 1 for disagreements that had BLAST score within 75% of the larger score, 0.01 for disagreements with very different BLAST scores, and 0.1 for situations where one of the BLAST searches did not yield any results. For cases where metabolites are linked to reactions but there is not a reference protein sequence available, a weight factor, *X*, is needed. We chose, *X*, such that: i) X=0.01 when a metabolite is not in any reaction; ii) X=1.01 when a metabolite is in reaction missing a reference sequence; and iii) X=2.01 when a metabolite is in a reaction with a sequence. These arbitrary scores were selected solely to distinguish between different “agreement” states during the reciprocal BLAST. We did not observe much difference in the plurality of compound annotations depending on these weights (data not shown), however, they did have an impact on the number of annotations that agreed with KEGG and MetaCyc (Supplementary Figure 4). The most impactful weight appeared to be the “close reciprocal disagreement,” meaning that there was not an exact match in the bidirectional BLAST, but the e-scores were within the given threshold. If this weight was low (0.01) or high (2.0), there were fewer annotations that agreed with KEGG and MetaCyc.

The final MAGI-score $M = \mathrm{GM}\left(\left[C, H, \alpha, X\right]\right) / n^{L}$is a proxy for the probability that a gene and metabolite are associated. *M* is generated by calculating the geometric mean (*GM*) of the metabolite score (*C*), homology score (*H*), reciprocal agreement score (*α)* and weight factor (*X*), and whether or not the metabolite is present in a reaction ($n^L$) where *L* is the network level connecting the metabolite to a reaction (a proxy for the probability that a compound is involved in a reaction) and *n* is a penalty factor for the network level. Currently, n is equal to 4, but this parameter may change as the scoring function is optimized and more training data is acquired. We did not observe this penalty factor to greatly affect the number of gene annotations that agreed with KEGG or MetaCyc, though this did have a large impact on the number of features with multiple suggestions for compound identities; the higher the penalty factor, the lower the number of compound suggestions. MAGI often gives a high score to multiple metabolites, which is not surprising given the relatedness of many metabolites (e.g. isomers). Therefore, we recommend carefully considering the top scoring molecules and not assuming that the top ranked one is correct (Supplementary Figure 5). Additional benchmarking analysis shows agreement between KEGG, BioCyc and MAGI annotations for high MAGI homology scores (Supplementary Discussion and SI Figures 2 and 3). The geometric mean was used to account for the different scales of the individual scores, but weights may be applied to each individual score during the geometric mean calculation to further fine-tune the MAGI scoring process. We expect the weights to become further optimized as more results are processed through MAGI.

The final output is a table representing all unique metabolite-reaction-gene associations, their individual scores, and their integrated MAGI score (Supplementary Table 2). For scoring metabolite identities, a slice of this final output is created by retaining the top scoring metabolite-reaction-gene association for each unique metabolite structure; these can be mapped back onto the mass spectrometry results table to aid the identification of each mass spectrometry feature. For assessing gene functions, another slice of this final output is created by retaining the top scoring metabolite-reaction-gene association for each unique gene-reaction pair. For a typical bacterial genome of ~ 6000 genes and a metabolites file of ~ 6000 compounds, the MAGI calculation performed via the web service at https://magi.nersc.gov/ should take about thirty minutes to complete. While MAGI can provide valuable insights into primary metabolism, these reactions tend to be better characterized and therefore a particularly important application of MAGI is for secondary metabolite pathways.

**Data Availability**

All source code is available at https://github.com/biorack/magi, and the *S. coelicolor* mass spectrometry data (.mzML files) and MIDAS results (metabolite_0ae82b08.csv) can be found here: https://magi.nersc.gov/jobs/?id=0ae82b08-b2a3-40d8-bb9a-e64b567cacd2.

**Application Availability and Usage**

Potential MAGI users may use the application on their personal computers by downloading the source code from the GitHub repository, or may upload their data files to the web service. In order to use MAGI, users must provide at least one of the following: a FASTA file of genes they wish to be associated to reactions, and/or a metabolites file they wish to associate to reactions. The metabolites file should be in a table file format (*e.g.* .csv, .tsv, Excel), and must have a column named “original_compound” that describes the InChI Key for each metabolite of interest. If both FASTA and metabolite files are provided, then  associations between genes and metabolites will be made as well.

## Results and Discussion

**Improved metabolite identification for metabolomics.** To examine how MAGI uses genomic information to filter and score possible metabolite identities from a metabolomics experiment, sequencing and metabolomics data were obtained for *S. coelicolor*. After processing the raw LCMS data to find chromatograms and peaks, 878 features with a unique *m/z* and retention time were found in the dataset. After neutralizing the *m/z* values, accurate mass searching, and conducting MS/MS fragmentation pattern analysis, 6,604 unique metabolite structures were tentatively associated with these features (Supplementary Table 3). This means on average there were almost 8 candidate structures for each feature. For a candidate structure to be associated with a feature, it must have at least one matching fragmentation spectrum. As this is often the method for identifying metabolites, it highlights the problem in deconvolution of a signal to a specific chemical structure. 2,786 of these structures were then linked to a total of 10,265 reactions either directly or via the chemical similarity network, and the reactions were associated with 3,181 (out of 8,210) *S. coelicolor* genes by homology. Finally, a MAGI score was calculated for each metabolite-reaction-gene association (Supplementary Table 4).

An example that illustrates MAGI’s utility in determining the most likely correct metabolite identification is the feature putatively identified as 1,4-dihydroxy-6-naphthoic acid.  Here, a feature with an *m/z* of 203.0345 was observed. This feature was associated with the chemical formula $C_{11}H_8O_4$, which could be derived from 16 unique chemical structures in the metabolite database (Supplementary Table 5). Mass fragmentation spectra were collected for this feature and analyzed using MIDAS[7], a tool that scores the observed fragmentation spectrum against its database of *in-silico* fragmentation trees for the 16 potential structures.  Based only on the MIDAS metabolite score, the top scoring structure was 5,6-dihydroxy-2-methylnaphthalene-1,4-dione. However, after calculating the MAGI scores, a different metabolite received the highest score. Of the 16 potential metabolites, only 1,4-dihydroxy-6-naphthoic acid was in a reaction that had a perfect match to genes in *S. coelicolor* (an E-value of 0.0 to *SCO4326*; Table 1). This metabolite is a known intermediate in an alternative menaquinone biosynthesis pathway discovered in *S. coelicolor*[41, 42], making it much more likely to be a metabolite detected from the metabolome of *S. coelicolor* as opposed to the metabolite found by looking at mass fragmentation alone.

**Metabolomics-driven gene annotations.** MAGI keeps the biochemical potential of an organism unconstrained by considering a plurality of probable gene product functions. One effect of this is that more reactions are associated with genes than other services (Figure 2A). Because reactions are the pivotal link between metabolites and genes, this allows integration of a larger fraction of a metabolomics dataset with genes. Furthermore, MAGI associates many genes that are not annotated using traditional approaches with at least one reaction (Figure 2B). Out of a total of 8,210 predicted coding sequences in *S. coelicolor*, KEGG and BioCyc have one or more reactions associated with 1,106 and 1,294 genes, respectively. On the other hand, MAGI associated 5,209 genes with one or more reactions, out of which 3,719 genes had no reaction associated with them in either KEGG or BioCyc (Figure 2B). Of these 3,719 genes, 1,883 were linked to at least one metabolite in the metabolomics data (Supplementary Table 4). Certainly, not all MAGI gene-reaction associations are correct, however, this does provide many testable hypotheses that give footholds to discover new biochemistry As can be seen in Figure 2C, many of these new gene-reaction associations have high scores, indicating a likely connection.

**Validation of gene-metabolite integration in pathways.** One of the most well-known biosynthetic pathways in *S. coelicolor* is the pathway to synthesize the pigmented antibiotic actinorhodin[35]. We examined the MAGI results involving the metabolites and genes of actinorhodin biosynthesis as a proof-of-principle that MAGI successfully integrates metabolites and genes, and that these results can be mapped onto a reaction network. Actinorhodin and all of its detected intermediates were correctly identified and accurately mapped to the correct genes (Figure 3A), despite some intermediates having several plausible metabolite identities (Supplementary Table 6). Notably, KEGG did not annotate the majority of actinorhodin biosynthesis genes, and the one gene that it did annotate was incorrect (Table 1).

In another example, we examined the menaquinone biosynthesis pathway, which is essential for respiration in bacteria[43] and thus should be included in every metabolic reconstruction for organisms that produce menaquinone. An alternative menaquinone biosynthesis pathway was recently discovered and validated in *S. coelicolor*[41, 42], serving as another proof-of-principle exercise for assessing the MAGI platform. MAGI linked 4 of 7 intermediate metabolites of the pathway to the appropriate genes (Figure 3B, Supplementary Table 7). Interestingly, while KEGG accurately assigned reactions to all but one of the genes in this biosynthetic pathway, BioCyC had vague textual annotations and no reactions (Table 1). Therefore, a metabolomics tool that relies on BioCyc model for *S. coelicolor* would be unable to integrate any of these metabolites with genes for the purpose of either improved metabolite identifications or gene annotations.

**Correction of annotation errors.** Gene annotation pipelines are notoriously error-prone[44] and yield inconsistent results based on the bioinformatic analyses used: the database used for homology searches, and what kind of additional data (*e.g.* PFams, genetic neighborhoods, and literature mining) are incorporated into the annotation algorithm or not (see Table 1 for some examples). For example, the undecylprodigiosin synthase gene is known[45], yet was incorrectly annotated in the KEGG genome annotation for *S. coelicolor*. KEGG annotated this gene as “PEP utilizing enzyme” with an EC number of 2.7.9.2 (pyruvate, water phosphotransferase with paired electron acceptors). This error is notable because the undecylprodigiosin synthase reaction has an EC number of 6.4.1.-: ligases that form carbon-carbon bonds. On the other hand, BioCyc correctly annotates *SCO5896* as undecylprodigiosin synthase, presumably using manual curation or a thorough literature-searching algorithm.

MAGI used metabolomics data to score the possible gene annotations for *SCO5896* in addition to homology scoring (*i.e.* E-value). In the absence of metabolomics data, MAGI initially associated the *SCO5896* gene sequence with the prodigiosin synthase and norprodigiosin synthase reactions via BLAST searches against the MAGI reaction reference sequence database (Figure 4). Metabolomics analysis revealed that the feature with an *m/z* of 392.2720 could potentially be undecylprodigiosin, which MAGI associated with only the undecylprodigiosin synthase reaction (Figure 4). Because this reaction does not have a reference sequence in our database, it could not be queried against the *S. coelicolor* genome. However, the chemical network revealed that prodigiosin is a similar metabolite that is in a reaction that does have a reference sequence (Figure 4). When the prodigiosin synthase reaction’s reference sequence was queried against the *S. coelicolor* genome, the top hit was *SCO5896*, thus making a reciprocal connection between the mass spectrometry feature and gene via the prodigiosin synthase reaction (Figure 4).

**Making nonexistent or vague annotations specific.** The vast majority of sequenced genes have no discrete functional predictions, preventing the in-depth understanding of metabolic processes of most organisms. *S. coelicolor* is well known to produce several polyketides and is known to have the genetic potential to produce many more. The *SCO5315* gene product is WhiE, a known polyketide aromatase involved in the biosynthesis of a white pigment characteristic of *S. coelicolor*[46, 47]. KEGG and BioCyC textually annotated the gene as “aromatase” or “polyketide aromatase,” but neither links the gene to a discrete reaction. Although the text annotations are correct, the lack of a biochemical reaction prohibits the association of this gene with metabolites. On the other hand, MAGI was successfully able to associate *SCO5315* with an observed metabolite (20-carbon polyketide intermediate with an *m/z* of 401.0887) via a polyketide cyclization reaction with a MAGI consensus score of 4.59 (Table 1). While the physiological function of WhiE is to cyclize a 24-carbon polyketide intermediate, the enzyme has been shown to also catalyze the cyclization of similar polyketides with varying chain length, including the 20-carbon species observed in the metabolomics data presented here[48-50].

In another example where other annotation services were unable to assign any reactions to a gene product, MAGI associated *SCO7595* with the anhydro-NAM kinase reaction via the detected metabolite anhydro-N-acetylmuramic acid (anhydro-NAM) (*m/z* 274.0941) (Table 1). Anhydro-NAM is an intermediate in bacterial cell wall recycling, a critically important and significant metabolic process in actively growing bacterial cells; *E. coli* and other bacteria were observed to recycle roughly half of cell wall components per generation[51, 52]. MAGI also associated anhydro-NAM to *SCO6300* via an acetylhexosaminidase reaction (Table 1) that produces the metabolite. KEGG and RAST both annotate this gene to be acetylhexosaminidase with a total of 5 possible reactions, but none involve anhydro-NAM (Table 1). The detection of anhydro-NAM may be considered orthogonal experimental evidence to indicate that *SCO6300* can act on N-acetyl-β-D-glucosamine-anhydro-NAM along with the other acetylhexosamines predicted by KEGG and RAST, forming an early stage in anhydromurpoeptide recycling. In the absence of MAGI, a researcher may have been able to manually curate a metabolic model by manually assessing the text annotations and adding reactions to the model, but the MAGI framework not only makes this process easier, it also connects an experimental observation that supports the predicted function of the gene.

**Potential for making novel annotations.** In addition to these few examples, there are hundreds more gene-reaction-metabolite associations that could be used to strengthen, validate, or correct existing annotations from KEGG or BioCyc, as well as discover new annotations through experimentation. These MAGI associations can be sorted by their MAGI score to generate a ranked list of candidate genes and gene functions, with optional hierarchical grouping and filtering of the list by homology, metabolite, chemical network, and/or reciprocal score. For example, of the 1,883 *S. coelicolor* genes that were uniquely linked to a metabolite via a reaction by MAGI, roughly one-third were connected directly to a metabolite; that is, the chemical similarity network was not used to expand reaction space (Figure 5A and Figure 2C teal markers). Furthermore, one-third of these genes had perfect reciprocal agreement between the metabolite-to-gene and gene-to-metabolite search directions (Figure 5B and Figure 2C teal circles). These 190 genes can be further separated or binned based on their homology score or MAGI score (Figure 5C), resulting in an actionable number of high-priority and high-strength novel gene function hypotheses to test in future studies.

**Limitations of this study.** In this study, we show that MAGI produces plausible associations between genes and metabolites from *Streptomyces coelicolor*. Since the associations shown in this paper are judged by manual inspection, there are not enough validated links to compute a reliable false discovery rate or applicability to other systems. Therefore an important future work will be to broadly apply MAGI across many organisms and evaluate the generality of this approach. This will ensure that the parameters used are not over fit specifically to *Streptomyces coelicolor*. In addition, given the paucity of direct biochemical validations of gene functions, it will likely be necessary to integrate MAGI with high throughput mutagenesis studies to accurately determine false discovery rates.  Lastly, more unique metabolites can be observed by combining data collected from polar and lipid fractions of metabolites along with combining positive and negative ionization modes. The results here are based on measured signals from a small subset of the Streptomyces coelicolor metabolome.

## Conclusion

In this work we describe MAGI, a method for integrating metabolomics observations with genomic predictions to help overcome the limitations of each and strengthen the biological conclusions made by both. Using *Streptomyces coelicolor* as a test case, we find that this method can help strengthen metabolite identifications, suggests specific biochemical predictions about genes that may otherwise be ambiguous, and suggests new biochemistry via the chemical network. It will be important to also evaluate this approach for diverse organisms to determine the generality of the method.  In order to facilitate broad usage by the academic community, we provide MAGI through the National Energy Research Scientific Computing Center (NERSC) at https://magi.nersc.gov, where users can upload their own metabolite and FASTA files for analysis through MAGI.

## Acknowledgements

This work was supported by the U.S. Department of Energy Office of Science by the Ecosystems and Networks Integrated with Genes and Molecular Assemblies (ENIGMA) Program, the U.S. Department of Energy Joint Genome Institute (JGI), and the National Energy Research Scientific Computing Center (NERSC) – a DOE Office of Science User Facility – all under Contract No. DE-AC02-05CH11231.

## References

1. Liu, X. J., and Locasale, J. W. (2017) Metabolomics: A Primer, *Trends in Biochemical Sciences 42*, 274-284.
2. Creek, D. J., Dunn, W. B., Fiehn, O., Griffin, J. L., Hall, R. D., Lei, Z. T., Mistrik, R., Neumann, S., Schymanski, E. L., Sumner, L. W., Trengove, R., and Wolfender, J. L. (2014) Metabolite identification: are you sure? And how do your peers gauge your confidence?, *Metabolomics 10*, 350-353.
3. Wolfender, J. L., Marti, G., Thomas, A., and Bertrand, S. (2015) Current approaches and challenges for the metabolite profiling of complex natural extracts, *J Chromatogr A 1382*, 136-164.
4. Vaniya, A., and Fiehn, O. (2015) Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics, *Trac-Trend Anal Chem 69*, 52-61.
5. Smith, C. A., O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R., and Siuzdak, G. (2005) METLIN: a metabolite mass spectral database, *Ther Drug Monit 27*, 747-751.
6. Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M. Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, K., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., and Nishioka, T. (2010) MassBank: a public repository for sharing mass spectral data for life sciences, *J Mass Spectrom 45*, 703-714.
7. Wang, Y., Kora, G., Bowen, B. P., and Pan, C. (2014) MIDAS: a database-searching algorithm for metabolite identification in metabolomics, *Anal Chem 86*, 9496-9503.
8. Wolf, S., Schmidt, S., Muller-Hannemann, M., and Neumann, S. (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra, *BMC Bioinformatics 11*, 148.
9. Allen, F., Greiner, R., and Wishart, D. (2015) Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification, *Metabolomics 11*, 98-110.
10. Ridder, L., van der Hooft, J. J. J., Verhoeven, S., de Vos, R. C. H., Bino, R. J., and Vervoort, J. (2013) Automatic Chemical Structure Annotation of an LC-MSn Based Metabolic Profile from Green Tea, *Analytical Chemistry 85*, 6033-6040.
11. Duhrkop, K., Shen, H. B., Meusel, M., Rousu, J., and Bocker, S. (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID, *Proceedings of the National Academy of Sciences of the United States of America 112*, 12580-12585.
12. Dhanasekaran, A. R., Pearson, J. L., Ganesan, B., and Weimer, B. C. (2015) Metabolome searcher: a high throughput tool for metabolite identification and metabolic pathway mapping directly from mass spectrometry and using genome restriction, *Bmc Bioinformatics 16*.

13. Li, S. Z., Park, Y., Duraisingham, S., Strobel, F. H., Khan, N., Soltow, Q. A., Jones, D. P., and Pulendran, B. (2013) Predicting Network Activity from High Throughput Metabolomics, *Plos Computational Biology 9*.

14. 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 Research 44*, D471-D480.

15. Morgat, A., Lombardot, T., Axelsen, K. B., Aimo, L., Niknejad, A., Hyka-Nouspikel, N., Coudert, E., Pozzato, M., Pagni, M., Moretti, S., Rosanoff, S., Onwubiko, J., Bougueleret, L., Xenarios, I., Redaschi, N., and Bridge, A. (2017) Updates in Rhea - an expert curated resource of biochemical reactions, *Nucleic Acids Research 45*, D415-D418.

16. Yang, J. Y., Sanchez, L. M., Rath, C. M., Liu, X. T., Boudreau, P. D., Bruns, N., Glukhov, E., Wodtke, A., de Felicio, R., Fenner, A., Wong, W. R., Linington, R. G., Zhang, L. X., Debonsi, H. M., Gerwick, W. H., and Dorrestein, P. C. (2013) Molecular Networking as a Dereplication Strategy, *J Nat Prod 76*, 1686-1699.

17. Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A., and Hatzimanikatis, V. (2016) ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies, *Acs Synthetic Biology 5*, 1155-1166.

18. Hatzimanikatis, V., Li, C. H., Ionita, J. A., Henry, C. S., Jankowski, M. D., and Broadbelt, L. J. (2005) Exploring the diversity of complex metabolic networks, *Bioinformatics 21*, 1603-1609.

19. Li, C. H., Henry, C. S., Jankowski, M. D., Ionita, J. A., Hatzimanikatis, V., and Broadbelt, L. J. (2004) Computational discovery of biochemical routes to specialty chemicals, *Chem Eng Sci 59*, 5051-5060.

20. Li, L., Li, R., Zhou, J., Zuniga, A., Stanislaus, A. E., Wu, Y., Huan, T., Zheng, J., Shi, Y., Wishart, D. S., and Lin, G. (2013) MyCompoundID: using an evidence-based metabolome library for metabolite identification, *Anal Chem 85*, 3401-3408.

21. Huan, T., Tang, C., Li, R., Shi, Y., Lin, G., and Li, L. (2015) MyCompoundID MS/MS Search: Metabolite Identification Using a Library of Predicted Fragment-Ion-Spectra of 383,830 Possible Human Metabolites, *Anal Chem 87*, 10619-10626.

22. Menikarachchi, L. C., Hill, D. W., Hamdalla, M. A., Mandoiu, II, and Grant, D. F. (2013) In silico enzymatic synthesis of a 400,000 compound biochemical database for nontargeted metabolomics, *J Chem Inf Model 53*, 2483-2492.

23. Jeffryes, J. G., Colastani, R. L., Elbadawi-Sidhu, M., Kind, T., Niehaus, T. D., Broadbelt, L. J., Hanson, A. D., Fiehn, O., Tyo, K. E., and Henry, C. S. (2015) MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics, *J Cheminform 7*, 44.

24. Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A., and Hatzimanikatis, V. (2016) ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies, *ACS Synth Biol 5*, 1155-1166.

25. Duigou, T., du Lac, M., Carbonell, P., and Faulon, J. L. (2018) RetroRules: a database of reaction rules for engineering biology, *Nucleic Acids Res*.

26. Kumar, A., Wang, L., Ng, C. Y., and Maranas, C. D. (2018) Pathway design using de novo steps through uncharted biochemical spaces, *Nat Commun 9*, 184.

27. Hattori, M., Tanaka, N., Kanehisa, M., and Goto, S. (2010) SIMCOMP/SUBCOMP: chemical structure search servers for network analyses, *Nucleic Acids Res 38*, W652-656.

28. Johnston, C. W., Skinnider, M. A., Wyatt, M. A., Li, X., Ranieri, M. R., Yang, L., Zechel, D. L., Ma, B., and Magarvey, N. A. (2015) An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products, *Nat Commun 6*, 8421.

29. Medema, M. H., Paalvast, Y., Nguyen, D. D., Melnik, A., Dorrestein, P. C., Takano, E., and Breitling, R. (2014) Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products, *PLoS Comput Biol 10*, e1003822.

30. Sevin, D. C., Fuhrer, T., Zamboni, N., and Sauer, U. (2017) Nontargeted in vitro metabolomics for high-throughput identification of novel enzymes in Escherichia coli, *Nat Methods 14*, 187-194.

31. Lamb, D. C., Guengerich, F. P., Kelly, S. L., and Waterman, M. R. (2006) Exploiting Streptomyces coelicolor A3(2) P450s as a model for application in drug discovery, *Expert Opin Drug Metab Toxicol 2*, 27-40.

32. Chater, K. F. (2016) Recent advances in understanding Streptomyces, *F1000Res 5*, 2795.

33. Worthen, D. B. (2008) Streptomyces in Nature and Medicine: The Antibiotic Makers, *Journal of the History of Medicine and Allied Sciences 63*, 273-274.

34. Craney, A., Ahmed, S., and Nodwell, J. (2013) Towards a new science of secondary metabolism, *J Antibiot (Tokyo) 66*, 387-400.

35. Craney, A., Ahmed, S., and Nodwell, J. (2013) Towards a new science of secondary metabolism, *J Antibiot 66*, 387-400.

36. Pluskal, T., Castillo, S., Villar-Briones, A., and Oresic, M. (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, *BMC Bioinformatics 11*, 395.

37. Bowen, B. P., and Northen, T. R. (2010) Dealing with the unknown: metabolomics and metabolite atlases, *J Am Soc Mass Spectrom 21*, 1471-1476.

38. Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M. (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways, *J Am Chem Soc 125*, 11853-11865.

39. 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*, W182-185.

40. Oellien, F., Cramer, J., Beyer, C., Ihlenfeldt, W. D., and Selzer, P. M. (2006) The impact of tautomer forms on pharmacophore-based virtual screening, *J Chem Inf Model 46*, 2342-2354.

41. Hiratsuka, T., Furihata, K., Ishikawa, J., Yamashita, H., Itoh, N., Seto, H., and Dairi, T. (2008) An alternative menaquinone biosynthetic pathway operating in microorganisms, *Science 321*, 1670-1673.

42. Mahanta, N., Fedoseyenko, D., Dairi, T., and Begley, T. P. (2013) Menaquinone Biosynthesis: Formation of Aminofutalosine Requires a Unique Radical SAM Enzyme, *Journal of the American Chemical Society 135*, 15318-15321.

43. Nowicka, B., and Kruk, J. (2010) Occurrence, biosynthesis and function of isoprenoid quinones, *Bba-Bioenergetics 1797*, 1587-1605.

44. Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009) Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies, *Plos Computational Biology 5*.

45. Haynes, S. W., Sydor, P. K., Stanley, A. E., Song, L. J., and Challis, G. L. (2008) Role and substrate specificity of the Streptomyces coelicolor RedH enzyme in undecylprodiginine biosynthesis, *Chem Commun*, 1865-1867.

46. Shen, Y. M., Yoon, P., Yu, T. W., Floss, H. G., Hopwood, D., and Moore, B. S. (1999) Ectopic expression of the minimal whiE polyketide synthase generates a library of aromatic polyketides of diverse sizes and shapes, *Proceedings of the National Academy of Sciences of the United States of America 96*, 3622-3627.

47. Yu, T. W., Shen, Y. M., McDaniel, R., Floss, H. G., Khosla, C., Hopwood, D. A., and Moore, B. S. (1998) Engineered biosynthesis of novel polyketides from Streptomyces spore pigment polyketide synthases, *Journal of the American Chemical Society 120*, 7749-7759.

48. Alvarez, M. A., Fu, H., Khosla, C., Hopwood, D. A., and Bailey, J. E. (1996) Engineered biosynthesis of novel polyketides: Properties of the whiE aromatase/cyclase, *Nature Biotechnology 14*, 335-338.

49. Mcdaniel, R., Hutchinson, C. R., and Khosla, C. (1995) Engineered Biosynthesis of Novel Polyketides - Analysis of Tcmn Function in Tetracenomycin Biosynthesis, *Journal of the American Chemical Society 117*, 6805-6810.

50. Ames, B. D., Korman, T. P., Zhang, W. J., Smith, P., Vu, T., Tang, Y., and Tsai, S. C. (2008) Crystal structure and functional analysis of tetracenomycin ARO/CYC: Implications for cyclization specificity of aromatic polyketides, *Proceedings of the National Academy of Sciences of the United States of America 105*, 5349-5354.

51. Park, J. T., and Uehara, T. (2008) How bacteria consume their own exoskeletons (Turnover and recycling of cell wall peptidoglycan), *Microbiology and Molecular Biology Reviews 72*, 211-227.

52. Johnson, J. W., Fisher, J. F., and Mobashery, S. (2013) Bacterial cell-wall recycling, *Ann Ny Acad Sci 1277*, 54-75.

53. Cooper, L. E., Fedoseyenko, D., Abdelwahed, S. H., Kim, S. H., Dairi, T., and Begley, T. P. (2013) In Vitro Reconstitution of the Radical S-Adenosylmethionine Enzyme MqnC Involved in the Biosynthesis of Futalosine-Derived Menaquinone, *Biochemistry 52*, 4592-4594.

54. Ichinose, K., Surti, C., Taguchi, T., Malpartida, F., Booker-Milburn, K. I., Stephenson, G. R., Ebizuka, Y., and Hopwood, D. A. (1999) Proof that the actVI genetic region of Streptomyces coelicolor A3(2) is involved in stereospecific

pyran ring formation in the biosynthesis of actinorhodin, *Bioorganic & Medicinal Chemistry Letters 9*, 395-400.

55. Taguchi, T., Itou, K., Ebizuka, Y., Malpartida, F., Hopwood, D. A., Surti, C. M., Booker-Milburn, K. I., Stephenson, G. R., and Ichinose, K. (2000) Chemical characterisation of disruptants of the Streptomyces coelicolor A3(2) actVI genes involved in actinorhodin biosynthesis, *J Antibiot 53*, 144-152.
56. Valton, J., Filisetti, L., Fontecave, M., and Niviere, V. (2004) A two-component flavin-dependent monooxygenase involved in actinorhodin biosynthesis in Streptomyces coelicolor, *Journal of Biological Chemistry 279*, 44362-44369.
57. Kendrew, S. G., Hopwood, D. A., and Marsh, E. N. G. (1997) Identification of a monooxygenase from Streptomyces coelicolor A3(2) involved in biosynthesis of actinorhodin: Purification and characterization of the recombinant enzyme, *Journal of Bacteriology 179*, 4305-4310.
58. Mcdaniel, R., Ebertkhosla, S., Fu, H., Hopwood, D. A., and Khosla, C. (1994) Engineered Biosynthesis of Novel Polyketides - Influence of a Downstream Enzyme on the Catalytic Specificity of a Minimal Aromatic Polyketide Synthase, *Proceedings of the National Academy of Sciences of the United States of America 91*, 11542-11546.

# Figures and Tables

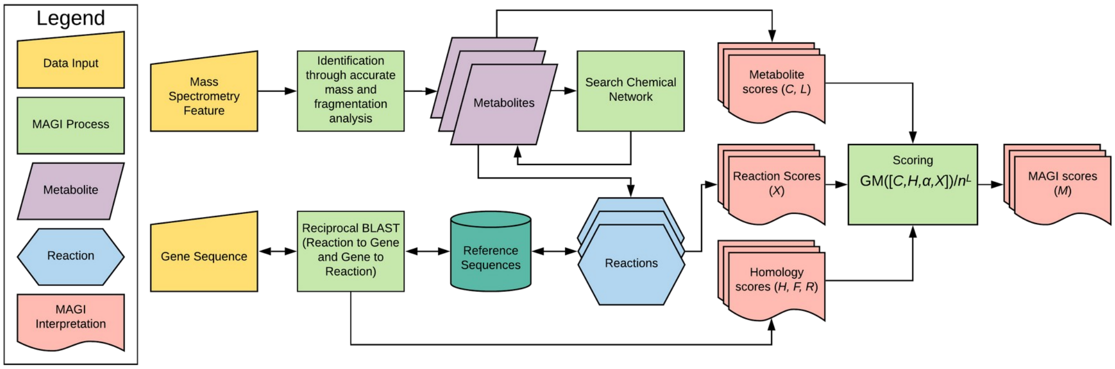

**Figure 1. MAGI workflow for consensus scoring**. Mass spectrometry features are connected to metabolites via methods such as accurate mass searching or fragmentation pattern matching. These metabolites are expanded to include similar metabolites by using the Chemical Network. These metabolites are then connected to reactions, which are reciprocally linked to input gene sequences via homology  (Reciprocal BLAST box). The metabolite, reaction, and homology scores generated throughout the MAGI process are integrated to form MAGI scores (Scoring box). For details on MAGI scores, see **Methods**.

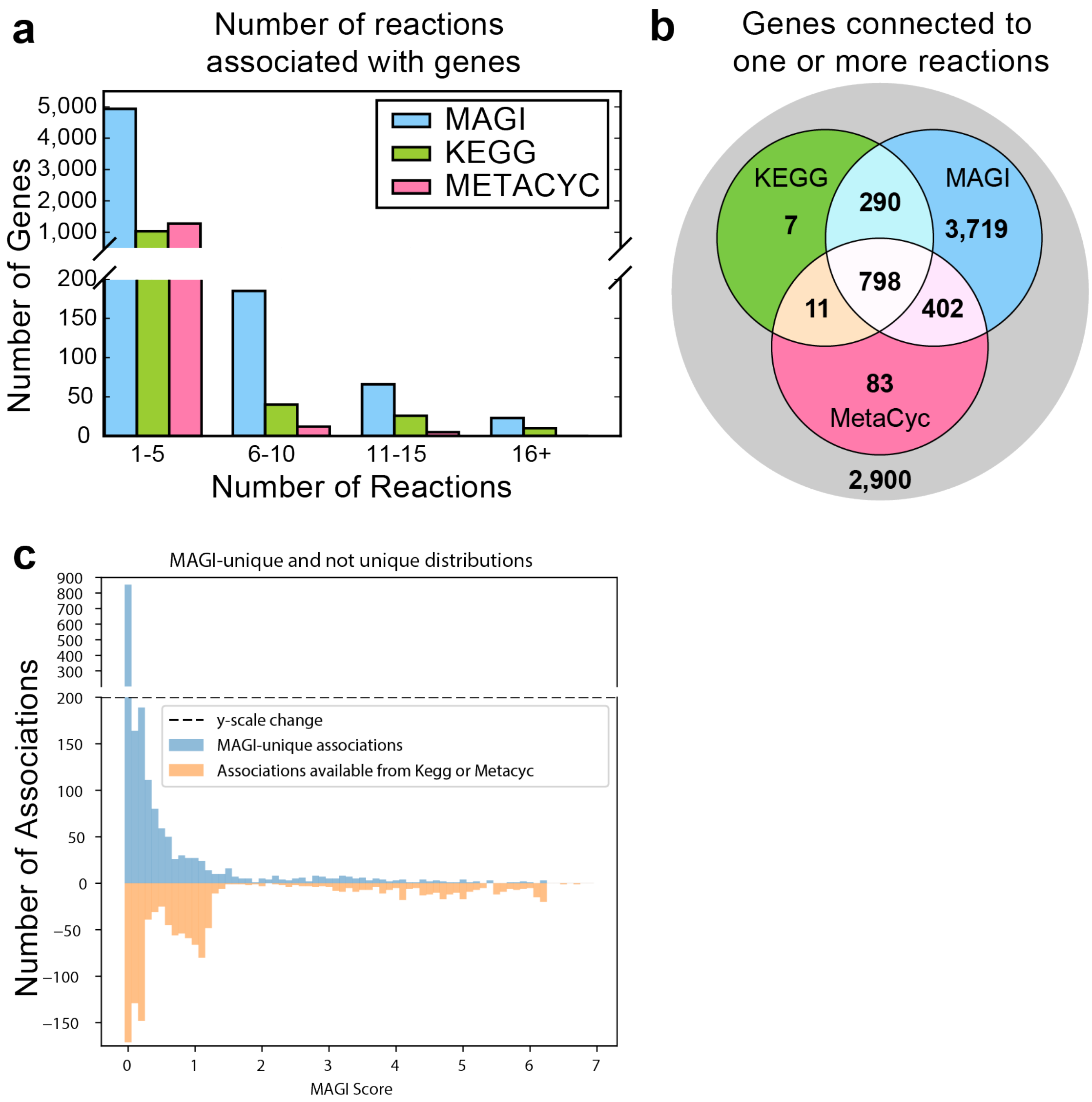

**Figure 2. MAGI associates more genes with reactions that can be ranked in *S. coelicolor*.** a) Number of reactions associated with each gene by MAGI, KEGG, and BioCyc. b) Venn diagram showing the genes connected to one or more reactions by MAGI, KEGG, and/or BioCyc. c) Distributions of the associations between a gene and a reaction for genes that have annotations in MetaCyc or Kegg (orange), or are unique to MAGI (blue), highlighting that there are several high-scoring MAGI associations for genes with no annotation.

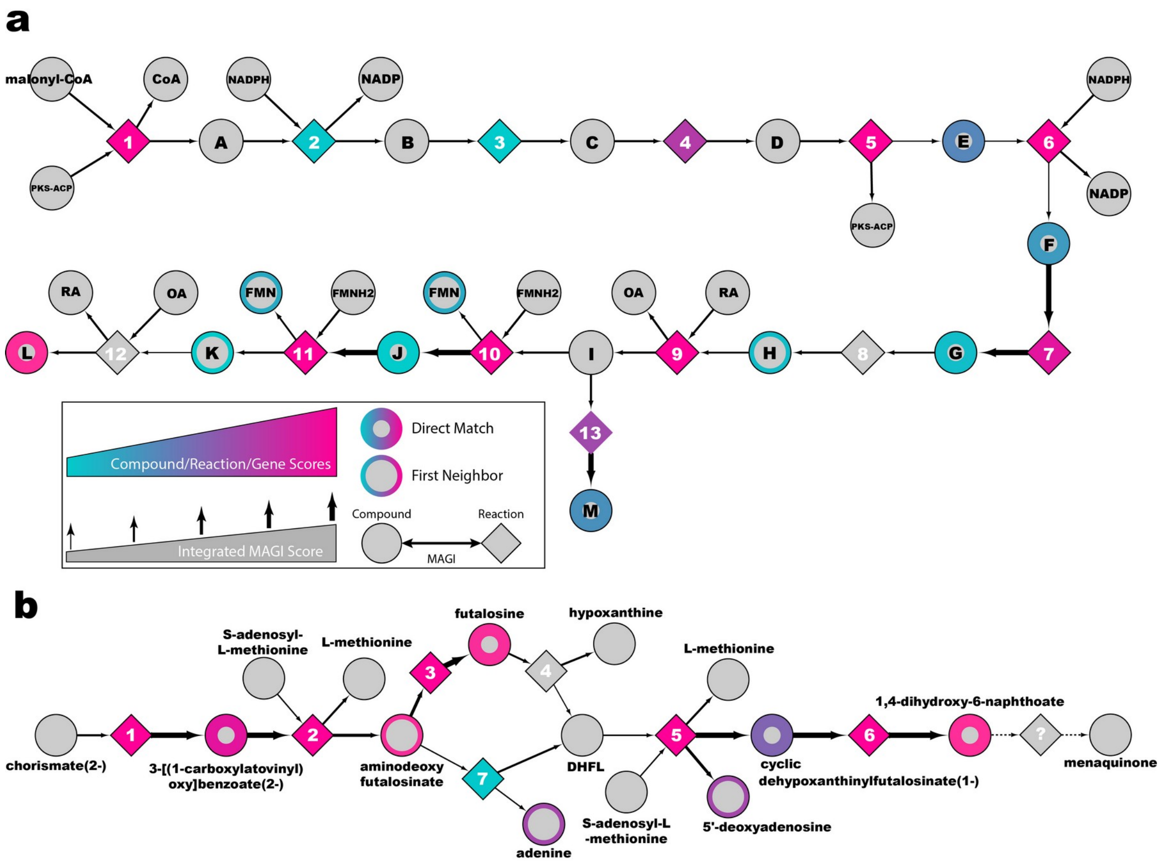

**Figure 3. Pathway views of MAGI results.** Metabolite, homology, and integrative MAGI scores throughout the (a) actinorhodin and (b) menaquinone biosynthesis pathways guides MAGI interpretations by visualizing results in a broader context. Circular nodes represent metabolites, diamond nodes represent reactions, and edges represent MAGI consensus scores. Border color of circular nodes corresponds to the MIDAS metabolite score, and border width corresponds to the chemical network level searched in MAGI. Fill color of diamond nodes correspond to the homology score. The line width of the edges corresponds to the MAGI score. Abbreviations and legends for metabolites and reactions are in supplementary table 8. The final step(s) in the menaquinone biosynthesis are currently not known and are represented by dashed edges and a “?” as the reaction.

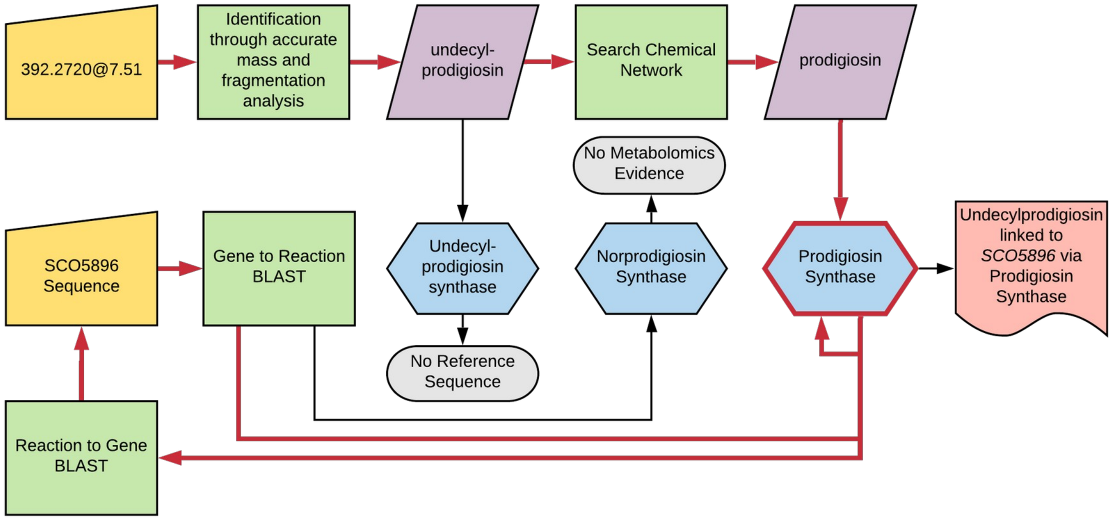

**Figure 4. Flowchart illustrating the key components of the MAGI algorithm and process for associating undecylprodigiosin with *SCO5896*.** In the upper half of the flowchart, the mass spectrometry feature with *m/z* 392.2720 at retention time 7.51 minutes was potentially identified to be undecylprodigiosin, which is in the undecylprodigiosin synthase reaction. This reaction has no reference sequence, so could not be directly connected to any *S. coelicolor* genes. Undecylprodigiosin was queried for similar metabolites in the chemical network, finding prodigiosin, which is in the prodigiosin synthase reaction. This reaction does have a reference sequence, which was used in a homology search against the *S. coelicolor* genome (Reaction to Gene BLAST), finding *SCO5896* as the top hit. In the lower half of the flowchart, the *SCO5896* gene sequence was queried against the entire MAGI reaction reference sequence database in a homology search (Gene to Reaction BLAST), finding the prodigiosin synthase and norprodigiosin synthase reactions. Norprodigiosin synthase did not have any metabolomics evidence, The metabolite-to-reaction and gene-to-reaction results were connected via the shared prodigiosin synthase reaction, effectively linking the feature 392.2720 to undecylprodigiosin and to *SCO5896*.

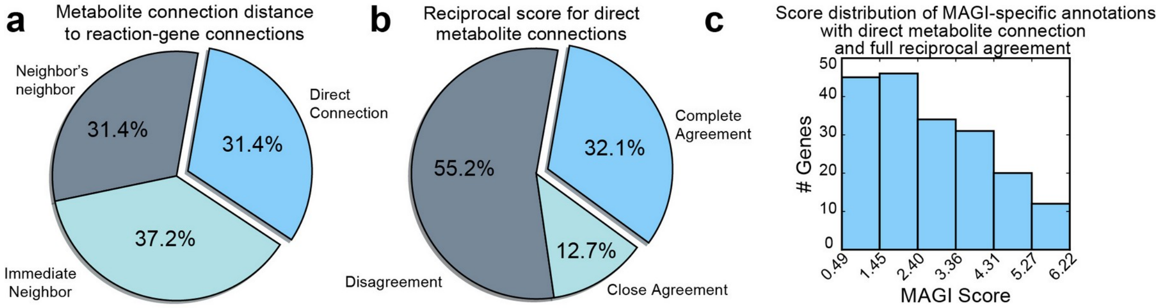

**Figure 5. Prioritization of MAGI gene function suggestions.** a) Of the 1,883 MAGI-specific gene-metabolite linkages (Figure 2C), 591 genes were associated with a reaction that was directly connected to an observed metabolite (*i.e.* the chemical similarity network was not used to link a metabolite to the reaction) (light blue). b) Of those, 190 genes had reciprocal agreement in bidirectional BLAST searches (light blue). c) Histogram of the top MAGI scores of the 190 genes from panel (b). Through this process an actionable number of high-priority and high-strength novel gene function hypotheses to test in future studies can be identified.

**Table 1. Comparison between MAGI, KEGG, and BioCyC annotations for *S. coelicolor* genes discussed in this study.**

| Gene | MAGI annotation (reaction) | MAGI score | Observed Metabolite Evidence | KEGG annotation (name) | KEGG Reaction Agreement with MAGI | BioCyc annotation (name) | BioCyc Reaction Agreement with MAGI |
|---|---|---|---|---|---|---|---|
| SCO4326 | RXN-10622 | 5.68 | Dihydroxy-naphthoate | 1,4-dihydroxy-6-naphthoate synthase | Agree | ORF | None |
| SCO4327 | RHEA:25907 | 5.16 | Futalosine | None | None | ORF | None |
| SCO4494 | RXN-15264 | 5.57 | Carboxy-vinyloxy-benzoic acid | Aminodeoxy-futalosine synthase | Agree | ORF | None |
| SCO4506 | RXN-12345 | 5.57 | Carboxy-vinyloxy-benzoic acid | chorismate dehydratase | Agree | ORF | None |
| SCO4550 | RXN-10620 | 5.03 | Cyclic-DHFL | cyclic dehypoxanthinyl futalosine synthase | Agree | ORF | None |
| SCO5074 | RXN1A0-6312 | 5.37 | Bicyclic intermediate F & (S)-Hemiketal | None | None | ActVI-ORF3 | Agree |
| SCO5075 | RXN1A0-6316 | 1.22 | Dihydro-kalafungin | None | None | ActVI-ORF4 | Agree |
| SCO5080 | RXN-18115 | 4.87 | DHK-red | 3-hydroxy-9,10-secoandrosta-1,3,5(10)-triene-9,17-dione monooxygenase [EC:1.14.14.12] | Disagree: R09819 | ActVA-ORF5 | Agree |
| SCO5081 | RXN1A0-6318 | 4.63 | Dihydro-kalafungin | None | None | ActVA-ORF6 | Agree |
| SCO5091 | RXN1A0-6307 | 5.95 | Bicyclic intermediate E | None | None | ActIV | Agree |
| SCO5315 | RXN-15413 | 4.58 | WhiE_20C_substrate | None | None | Polyketide aromatase | None |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| SCO5896 | RXN-15787* | 1.32 | Undecyl-prodigiosin | pyruvate, water dikinase | Disagree: R00199 | RedH | Agree* |
| SCO6300 | RXN0-5226 | 3.22 | Anhydro-NAM | beta-N-acetyl-hexosaminidase | Disagree: R00022, R05963, R07809, R07810, R10831 | hydrolase | None |
| SCO7595 | RHEA:24952 | 5.23 | Anhydro-NAM | anhydro-N-acetylmuramic acid kinase | None | ORF | None |

* Due to chemical network search, this reaction was listed as the prodigiosin synthase reaction but the metabolite connected to it was undecylprodigiosin, requiring manual interpretation to determine the actual reaction connected to the gene was undecylprodigiosin synthase.