PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Gene Prioritization by Compressive Data Fusion and Chaining

Marinka Žitnik[1], Edward A. Nam[2¤], Christopher Dinh[3], Adam Kuspa[2,3], Gad Shaulsky[2], Blaž Zupan[1,2]*

1 Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia, **2** Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, United States of America, **3** Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas, United States of America

¤ Current address: University of St. Thomas, Houston, Texas, United States of America
* blaz.zupan@fri.uni-lj.si

OPEN ACCESS

Citation: Žitnik M, Nam EA, Dinh C, Kuspa A, Shaulsky G, Zupan B (2015) Gene Prioritization by Compressive Data Fusion and Chaining. PLoS Comput Biol 11(10): e1004552. doi:10.1371/journal.pcbi.1004552

Editor: Niko Beerenwinkel, ETH Zurich, SWITZERLAND

Received: April 29, 2015

Accepted: September 12, 2015

Published: October 14, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. Project related code is available from GitHub repository (http://github.com/marinkaz/collage). The repository contains all data sets considered in the project.

Funding: This work was supported by the Dictyostelium Functional Genomics Program Project Grant from the NIH (P01 HD39691 - wet lab experiments), by grants from the ARRS (P2-0209, J2-5480 - development of the methods), by a grant from the EU (Health-F5-2010-242038 - development of the methods) and by Fulbright Scholarship (to BZ -

Abstract

Data integration procedures combine heterogeneous data sets into predictive models, but they are limited to data explicitly related to the target object type, such as genes. Collage is a new data fusion approach to gene prioritization. It considers data sets of various association levels with the prediction task, utilizes collective matrix factorization to compress the data, and chaining to relate different object types contained in a data compendium. Collage prioritizes genes based on their similarity to several seed genes. We tested Collage by prioritizing bacterial response genes in *Dictyostelium* as a novel model system for prokaryote-eukaryote interactions. Using 4 seed genes and 14 data sets, only one of which was directly related to the bacterial response, Collage proposed 8 candidate genes that were readily validated as necessary for the response of *Dictyostelium* to Gram-negative bacteria. These findings establish Collage as a method for inferring biological knowledge from the integration of heterogeneous and coarsely related data sets.

Author Summary

In everyday life, we make decisions by considering all the available information, and often find that inclusion of even seemingly circumstantial evidence provides an advantage. Our new computational method Collage prioritizes genes from a large collection of heterogeneous data. In a case study on social amoeba *Dictyostelium*, we started from four bacterial response genes and 14 different data sets ranging from gene expression to pathway and literature information. Collage proposed eight candidate genes that were tested in the wet laboratory. Mutations in all eight candidates reduced the ability of the amoebae to grow on Gram-negative bacteria. Furthermore, five out of the eight candidate genes were required for growth on Gram-negative bacteria but had no discernible effect on growth on Gram-positive bacteria. This is a remarkably accurate result since only about a hundred of the 12,000 *Dictyostelium* genes are estimated to be responsible for bacterial response.

research visits). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

In the natural sciences, incorporating all the data, especially circumstantial information, can be conceptually and computationally challenging. The difficulty stems from the heterogeneity and abundance of data sets. Consider a typical data analysis task in molecular biology: besides experimental data, such as levels of gene expression, there are plenty of other data sets at our disposal, such as protein-protein binding sites, genetic and metabolic pathways, functional annotations, genetic interactions, phenotype ontologies, diseases, drugs and their side effects. Intuitively, collective mining of all available information sources should improve accuracy of predictive modeling. However, the challenges are to integrate seemingly unrelated concepts from heterogeneous data sets [1] and fuse various data sets into a single predictive model.

Here we present a method called Collage that can consider a large number of potentially indirectly related data sets and use them for gene prioritization. Computational prediction of gene function is a formidable challenge. Given a small set of seed genes that are known to be responsible for a particular function, gene prioritization [2] aims to identify the most promising candidates for further studies. Present data integration approaches for gene prioritization can be divided into four groups: methods that consecutively filter one data set at a time [3]; methods that stitch together gene profiles from different data sources and then treat the stitched parts equally [4]; methods that use each data set separately to estimate the similarity of candidates to the seed genes and then fuse similarity scores through weighting [5–8, 8–12]; and methods that construct gene correlation networks independently from each data set and find genes that are similar to the seed genes in the composite network [13–17].

These approaches are limited to data that *explicitly* refer to genes. They cannot readily treat data that are relevant for gene prioritization but are provided in a non-gene data space, such as disease ontologies, phenotype classifications, drug interactions and annotations of small chemicals. A labor-intensive approach to consider data from non-gene space is feature engineering, which transforms circumstantial data into gene profiles. However, feature engineering is neither standardized nor effortless and is a bottleneck that prevents the implementation of truly large-scale data fusion for gene prioritization. As an alternative to gene-centric approaches, Collage represents a major advancement in (i) the breadth of data it can incorporate, (ii) the ease of data integration without complex feature engineering, (iii) the high prediction accuracy, (iv) the ability to retain the relational structure both within and between data sets during model inference and (v) the capacity to incorporate knowledge of data structure in model design.

We used Collage to solve a problem in an exciting and relatively new field of interest – the use of *Dictyostelium* as a model system to explore the interaction between eukaryotes and prokaryotes. *D. discoideum* is a free-living soil amoeba that feeds on bacteria. The amoebae eat both Gram-negative and Gram-positive bacteria, but they respond differently to bacteria from these two groups. Early studies have shown that mutations can impair the ability of the amoebae to grow on either Gram-positive or on Gram-negative bacteria [18]. Other studies have shown that the amoebae can serve as a model for the interaction between eukaryotes and prokaryotes, including pathogenesis [19–21]. This system is an important addition to the field because *Dictyostelium* is a very convenient model organism that offers a variety of experimental tools, including classical genetics and modern genomic approaches.

The interaction between *D. discoideum* and several Gram-positive and Gram-negative bacteria has recently been explored with genetic and genomic methods [22]. These studies revealed transcriptome-level responses to the two bacterial groups and discovered a handful of genes that are essential for growth of amoebae on bacteria. The genetic analysis suggested that one in a hundred of the 12,000 genes in the *D. discoideum* genome is required for bacterial

discrimination [22]. Identifying and characterizing these genes is a laborious task that requires several months of work per gene. We hypothesized that Collage could simplify this task by prioritizing genes and suggesting which ones should be tested by direct experiments.

Results

Compressive data fusion

Collage starts with a collection of data sets and can consider any kind of information (data tables, ontologies, associations, networks) that can be encoded in a matrix (S2 Fig). Each data set is viewed as a relation between two object types. For example, gene expression data relate gene names (columns) to experimental conditions (rows), where the entries represent transcript abundance. Literature annotation data relate research papers and their contents to annotation terms, where the entries are Boolean. Such data sets are abundant in the field of molecular biology and they report on dyadic relations that can be encoded in matrices. Matrix data representation is suitable for a wide range of data types, including tables, associations, ontologies and networks (S1 Fig). Whenever data sets share object types, we can connect them in a data fusion graph with object types as nodes and data matrices as edges. In the simplest data fusion graph shown in Fig 1a (top), node A may represent known genes in a certain genome and node B may denote various experimental conditions. A gene from A could be related to an experimental condition in B through a level of its mRNA abundance. Relationships between all genes and experimental conditions are represented in a data matrix that is placed on the edge A-B.

We model the system of data sets (Fig 1b) through data fusion by collective matrix factorization [23] (see also the tutorial provided in the S1 Text). Matrix factorization compresses the data matrices to a latent space and infers recipes to convert the latent representation back to the original data domain. Each data matrix is decomposed into a product of three low-dimensional latent matrices (S1 Fig): a "backbone matrix" encodes the relations between the latent components, and two "recipe matrices" transform the backbone matrix to the original space of the object types (Fig 1a). Data sets that are directly related and share a node in the fusion graph report on a common object type and hence use a common recipe matrix in their decomposition. Importantly, decomposition of any data set in the system depends on all other data sets according to a design of the fusion graph (Fig 1c). Sharing of recipe matrices ensures data fusion and allows Collage to incorporate knowledge about the relations between data sets.

Chaining of latent matrices

Collage profiles objects in the latent space of any other object type based on the connectivity in the data fusion graph. In the simplest scenario, where object types are adjacent, such as A and D in Fig 1c, Collage profiles objects of type A in the latent space of D by multiplying the recipe matrix of A by the backbone matrix A-D. The resulting profile matrix has objects of type A in rows and the latent components of type D in columns. The advantage of Collage over other gene prioritization tools is its ability to profile objects whose types are not direct neighbors in the fusion graph, such as A and C in Fig 1c. To profile objects of A in the latent space of C Collage starts with the recipe matrix of A and multiplies it by backbone matrices A-D, D-F and F-C on the path from A to C (Fig 1d). If A represented genes, D literature, F literature annotations and C chemical compounds, this procedure would yield profiles of genes in the latent space of chemical compounds. We refer to this technique as latent matrix chaining. It constructs dense profiles that include the most informative features obtained by collectively compressing data via matrix factorization. Intuitively, chaining is able to establish links between

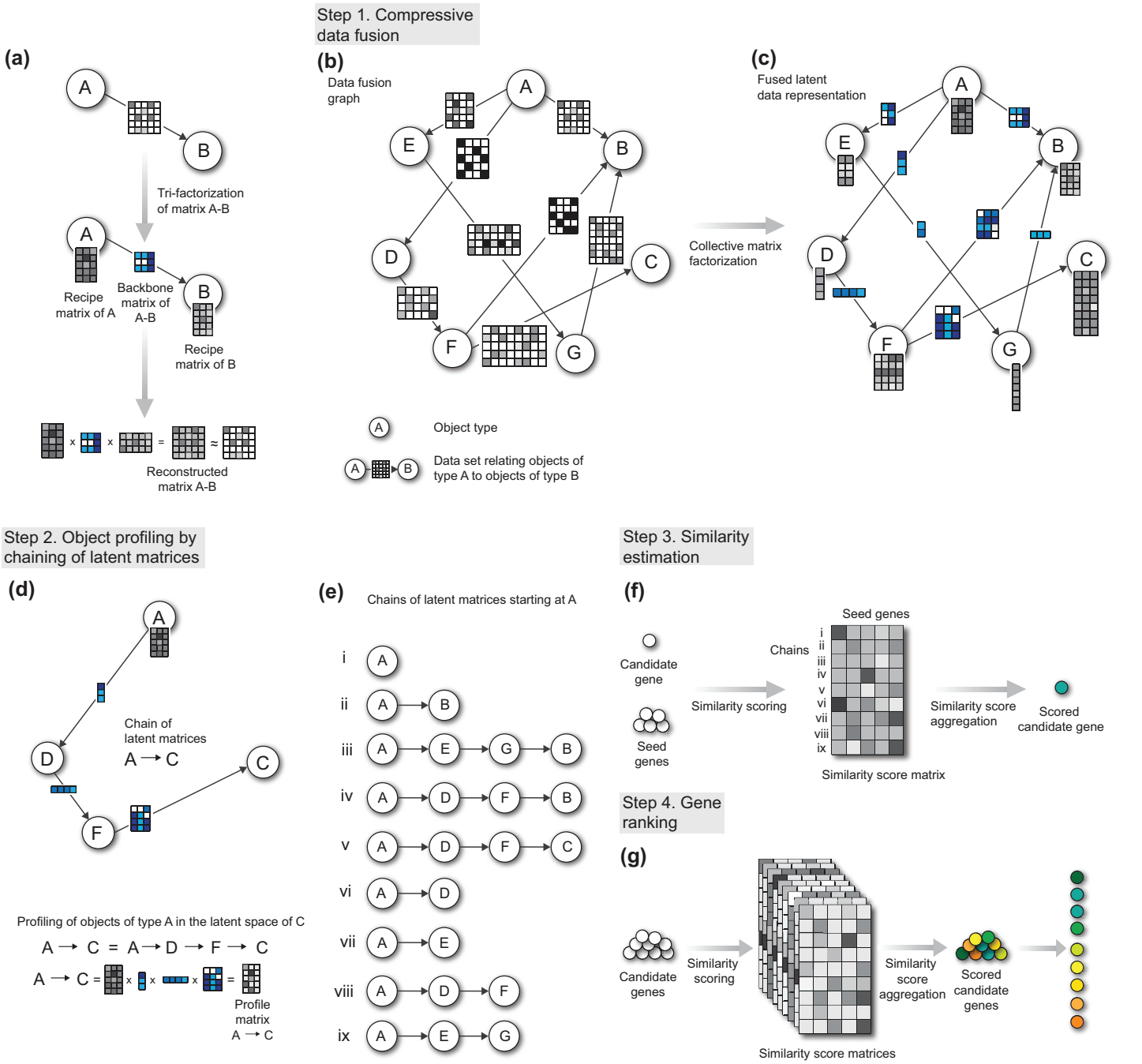

Fig 1. Overview of the Collage prioritization algorithm. (a) A data matrix in Collage relates two object types. We graphically represent this relation such that nodes A and B represent object types, and the directed edge A-B connects the two nodes with an associated data matrix. The matrix has objects of type A (e.g., genes) in the rows and objects of type B (e.g., experimental conditions) in the columns, as indicated by the edge directionality. Grey cells in the matrix represent quantitative measurements (e.g., mRNA transcript abundance), or binary memberships that relate objects in rows to objects in columns. Empty cells denote missing values. **(a, bottom)**. To model this relation, tri-factorization decomposes the data matrix into three smaller, low-dimensional latent matrices, whose product should well reconstruct the original matrix. Two latent "recipe matrices" map objects A and B into the latent space, and the remaining "backbone matrix" describes the relations in the latent space. In essence, the backbone matrix is a compressed version of the original data matrix. **(b)** Collage collectively models many data matrices that share object types. We organize the matrices in a data fusion graph. Object types are denoted as nodes (A to G), which may correspond to genes, ontology terms, diseases and patients, etc. **(c)** Instead of separately tri-factorizing each data matrix, Collage collectively factorizes all the matrices to a set of backbone matrices (edges, matrices in blue, one for each original data matrix) and recipe matrices (nodes, one for each object type), where the recipe matrices are shared across data sets that report on a common object type. **(d)** Collage chains latent matrices of the resulting factorized model to profile target objects (e.g., genes) in the latent space of any other object type. For example, the profiling of objects A in the latent space C is constructed by chaining that starts at node A and traverses the graph to node C through D and F. **(d, bottom)** Chaining multiplies the recipe

matrix A by the backbone matrices along the traversed path. **(e)** The A-to-C path in **(d)** is one of nine chains through which we can profile objects A in our exemplar data fusion graph. **(f)** The number of profile vectors for each object of type A corresponds to the number of chains. Collage compares the profiles of candidate genes to the profiles of the seed genes. Given a candidate gene, Collage records its rank correlation-based similarities in a similarity score matrix with seed genes in the columns and chained profiles in the rows. The final score estimates the similarity of a candidate to a set of seed genes and is obtained by summarizing the similarity score matrix with a single value (green circle) computed by a median-based L-estimator. **(g)** The similarity score of a gene is a proxy for its degree of involvement in the phenotype characterized by the set of seed genes. Hence the prioritization is defined by ranking the candidates according to their seed-similarity scores.

doi:10.1371/journal.pcbi.1004552.g001

genes and chemical compounds even though relationships between these object types are not available in input data in Fig 1b.

Gene prioritization

Collage prioritizes objects of the target object type (e.g., genes, node A in Fig 1) based on a small set of seed objects (previously characterized genes). For each target object, it constructs a set of profile matrices by considering all possible chains of latent matrices that start in the target node and end in any node that is reachable in data fusion graph (Fig 1e). A profile matrix corresponds to a particular latent matrix chain and encodes the latent space of the chain's last node. Each profile matrix is used to estimate the similarity between any two targets (genes) by comparing their respective profiles. Collages estimates the overall similarity between a candidate gene and the seed genes by aggregating similarity scores of the candidate gene across all profile matrices (Fig 1f). As a final step, Collage ranks all the genes based on their overall similarity with the seed genes (Fig 1g).

Bacterial response gene prioritization in *Dictyostelium*

Collage is agnostic to data types it can consider and can be applied to any collection of data sets and any phenotype of interest. We used Collage to find genes that affect *D. discoideum* growth on the Gram-negative bacteria *Klebsiella pneumoniae*. We started with four seed genes that have been previously identified in a genetic screen for *D. discoideum* mutants that fail to grow on Gram-negative bacteria (Table 1). We fused 14 publicly available data sets that were considered relevant to the problem. Collectively, these data sets describe relations between 10 object types (see data fusion graph in Fig 2). Our prioritization task was particularly challenging since there is not a lot of information about *Dictyostelium* in the literature and in public databases

Table 1. Seed *D. discoideum* genes used for Gram-negative bacterial response gene prioritization. Seed genes used for prioritization by Collage were selected based on the experiments published in [22].

Gene	DictyBase ID	Description
nip7	DDBG0295477	Ortholog of the conserved NIP7 nucleolar protein that is required for 60S ribosome subunit biogenesis; contains a PUA domain.
clkB	DDBG0278487	Similar to the cell division cycle 2-related protein kinase 7 (CRK7) and other cell division cycle 2-like protein kinases; belongs to the CMGC group of protein kinases.
spc3	DDBG0290851	Ortholog of the conserved microsomal signal peptidase 23 kDa subunit; the signal peptidase complex is a membrane-bound endoproteinase that removes signal peptides from nascent proteins as they are translocated into the lumen of the endoplasmic reticulum; contains a putative signal peptide.
alyL	DDBG0286229	Amoeba lysozyme family protein (aly), but divergent compared to *alyA-D*.

doi:10.1371/journal.pcbi.1004552.t001

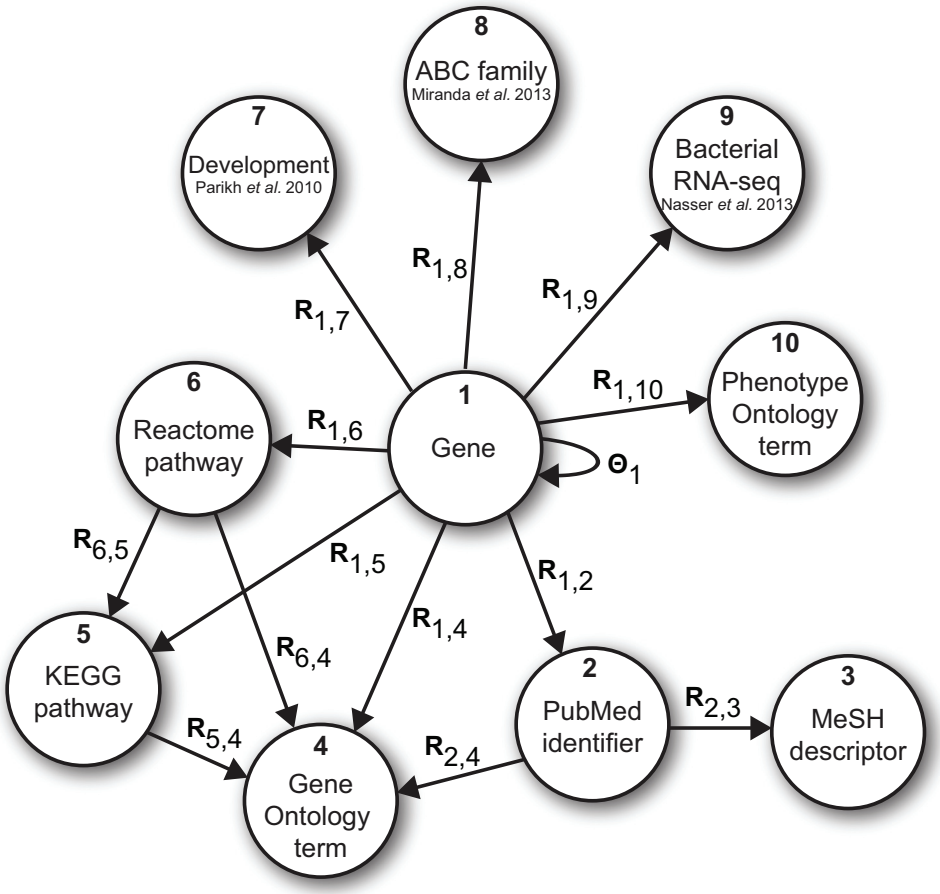

Fig 2. Data fusion graph for bacterial response gene prioritization in *Dictyostelium*. The graph shows the configuration of data sets, wherein nodes correspond to object types and edges denote data sets, each describing a relation between objects of two types. Collage considered 14 data sets (edges, represented by arrows) in this study describing the relations between 10 object types (nodes, represented by circles). The data sets included three whole-genome *D. discoideum* RNA-seq experiments [22, 24, 25] ($\mathbf{R}_{1,7}$, $\mathbf{R}_{1,8}$, $\mathbf{R}_{1,9}$), protein-protein interactions from the STRING database [26] ($\mathbf{\Theta}_1$), gene mentions in research articles ($\mathbf{R}_{1,2}$) and their Medical Subject Headings (MeSH) annotations ($\mathbf{R}_{2,3}$), pathway memberships from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [27] and Reactome [28] databases ($\mathbf{R}_{1,6}$, $\mathbf{R}_{1,5}$, $\mathbf{R}_{6,5}$), associations of genes to phenotypes from Phenotype Ontology [29] ($\mathbf{R}_{1,10}$), gene functions in Gene Ontology [30] ($\mathbf{R}_{1,4}$) and interrelatedness of Reactome and KEGG pathways and research literature with Gene Ontology terms ($\mathbf{R}_{6,4}$, $\mathbf{R}_{5,4}$, $\mathbf{R}_{2,4}$). See S4 Table for a detailed overview of considered data sets.

doi:10.1371/journal.pcbi.1004552.g002

and only one of the data sets (Fig 2, Bacterial RNA-seq, node 9) was directly related to bacterial response in *Dictyostelium*. Furthermore, the four seed genes, which were available to us at the beginning of this study, differ substantially in their data representation across data sets (S7 Fig). Collage ranked $\sim$ 12,000 genes from the *Dictyostelium* genome (S1 Table). The prioritized gene list was then filtered by the reported availability of *D. discoideum* gene knockout strains in the Dicty Stock Center (http://dictybase.org/StockCenter/StockCenter.html). We selected eight genes listed in Table 2 from the 30 top-ranked candidates (S2 and S3 Tables) for direct testing.

Table 2. Top-ranked candidate *D. discoideum* genes tested for Gram-negative bacterial response. The name of the candidate gene, its identifier and description from DictyBase are shown, together with the rank (out of all *D. discoideum* gene knockout strains available in the Dicty Stock Center) at which the candidate was prioritized. Collage prioritized genes by fusing data sets from the fusion graph shown in Fig 2.

Gene	DictyBase ID	Description	Rank position
cf50-1	DDBG0273175	Component of the counting factor complex, which includes CF60, CF50, CF45-1, and CtnA (countin).	1
smlA	DDBG0287587	Cytosolic protein present in vegetative and developing cells.	2
acbA	DDBG0270658	Precursor of SDF-2; similar to diazepam binding inhibitor; enriched in prespore cells.	3
abpC	DDBG0269100	120 kDa F-actin binding protein also often called filamin; involved in actin cytoskeleton organization, motility, sand development; enriched in prestalk cells.	6
pikB	DDBG0283081	Phosphatidylinositol kinase.	9
pikA	DDBG0278727	Phosphatidylinositol kinase.	11
pten	DDBG0286557	Phosphatase and tensin homolog.	15
modA	DDBG0269154	Protein post-translational modification mutant.	23

doi:10.1371/journal.pcbi.1004552.t002

Validation of top ranked candidate genes

To validate the selected candidate genes, we assessed growth of the *D. discoideum* knockout strains by making serial dilutions of the amoebae and co-culturing the cells with *K. pneumoniae* bacteria on nutrient agar. We observed a significant difference in the growth of all the mutants compared to the wild type AX4 (Fig 3). In this system, the bacteria grow faster than the amoebae so the first observation is the appearance of a thick opaque lawn of bacteria on the surface of the agar plate within 24 hours (not shown). Later on, as the amoebae eat the bacteria, they clear parts or all of the bacterial lawn, depending on their density and growth rate. When there are numerous, fast growing amoebae, we observe a cleared lawn (e.g. Fig 3, AX4, 10^4 cells, Day 2). When there are very few amoebae, we observe distinct plaques that appear as darker spots in the bacterial lawn (e.g. Fig 3, AX4 Day 3, 10^2 cells). When the bacteria are consumed, the amoebae starve, aggregate, and form developmental structures (Fig 3, AX4 Day 3, 10^4 cells). Cells that carry an inactivating mutation in the *tirA* gene (*tirA*$^-$ cells) exhibit impaired growth on *K. pneumoniae* [31]. We used these cells as a control in our assay and indeed they exhibited no clearing of the bacterial lawn when plated at the same initial density as the wild type cells (Fig 3, AX4 vs. *tirA*$^-$, Day 2, 10^4 cells). We note that *tirA*$^-$ cells can grow to some extent on *K. pneumoniae* bacteria under certain conditions, indicating that the growth phenotype is continuous even though many researchers tend to describe it as Boolean.

We tested the predictions made by Collage on eight genes–*acbA*, *smlA*, *pikA*, *pikB*, *pten*, *abpC*, *modA* and *cf50-1* (Table 2). In the case of *pikA* and *pikB* we used a double knockout strain because of previously reported overlap in the functions of these two genes [32]. Strikingly, when we assessed the ability of the mutant cells to grow on bacteria, they all exhibited varying degrees of growth defects compared to the equivalent wild type (AX4) control (Fig 3). Comparing only one condition, disruption of *acbA*, *abpC* and *modA* resulted in small individual plaques in the bacterial lawn but not complete clearing as observed in AX4 (Fig 3, black box, Day 2, 10^4 cells). In contrast, mutations in *smlA*, *pikA*/*pikB*, *pten*, and *cf50-1* caused phenotypes as severe as the loss of *tirA* with no clearing on Day 2 (Fig 3, black box, Day 2, 10^4 cells). Further distinction in the ability to grow on bacteria was revealed when the mutant cells were observed for an additional day. For example on Day 2, *pikA*$^-$/*pikB*$^-$ and *pten*$^-$ cells

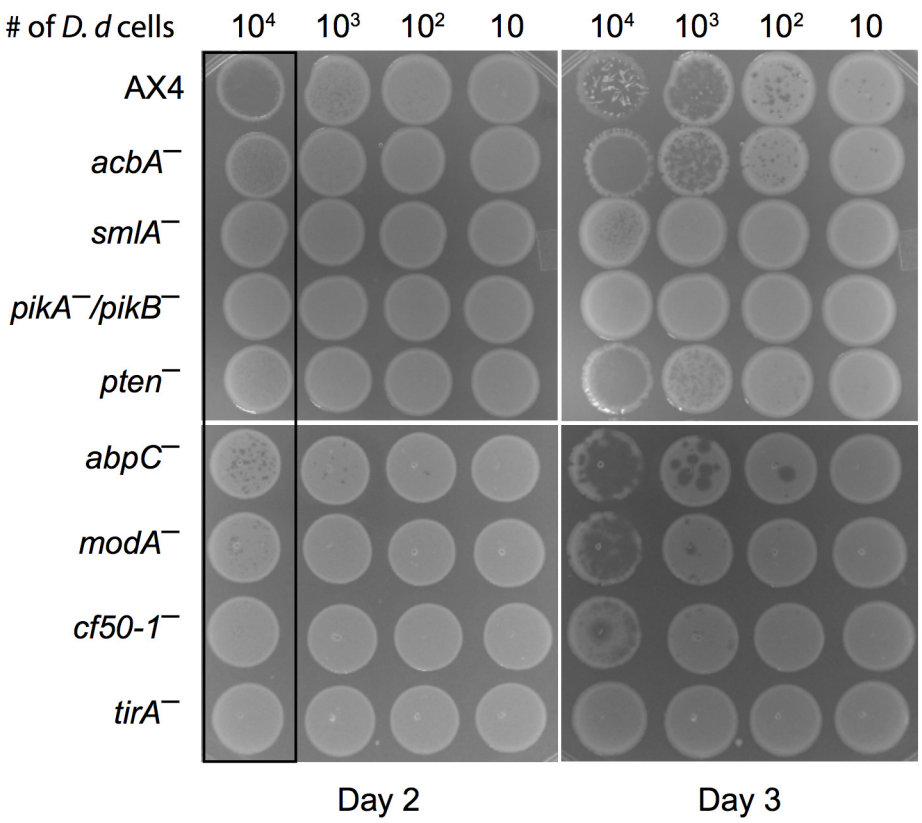

Fig 3. Experimental validation of top ranked candidate genes. Co-cultures of *D. discoideum* (*D.d.*) and bacteria were generated by serial dilutions of axenically grown *D. discoideum* amoebae with a large excess of *K. pneumoniae* bacteria such that the number of amoebae plated in each spot was between 10 and 10^4 as indicated above each column. The relevant genotypes of the amoebae strains are indicated on the left of each row. The co-cultures were plated on SM agar plates and incubated in a humid chamber at 22°C. Images were taken at 2 and 3 days after plating to show the progression of amoebae growth in time. The larger white opaque spots are lawns of the *K. pneumoniae* bacteria. Growth of the *D. discoideum* amoebae results in the formation of plaques within the opaque spots in cases of low amoebae cell density or clearing of much or all of the opaque spots in cases of high amoebae cell density. Upon complete clearing of the bacteria, the amoebae starve and begin to develop, producing white protruding multicellular structures within the lawn (e.g. AX4, 10^4 cells, day 3), which have no significance in this assay. Growth of the Collage-predicted knockout strains was compared to the wild type (AX4, top row) and to the most severe mutant available (*tirA*–, bottom row). Each experiment was performed in duplicate. Representative images of three independent experiments are shown.

doi:10.1371/journal.pcbi.1004552.g003

exhibited similar growth defects, but by Day 3, the loss of *pten* did not hinder growth on bacteria as much as the loss of *pikA* and *pikB* (Fig 3).

The seed genes we selected are required for growth on Gram-negative bacteria but dispensable for growth on Gram-positive bacteria [22]. This information was not included explicitly in our Collage analysis, but it was interesting to test the effect of the eight validated genes on growth on Gram-positive bacteria as well. We therefore plated the mutant strains on *Bacillus subtilis* bacteria and tested their growth. The wild type (AX4) control grew well, as did the *tirA* mutant, thus validating the assay. Disruption of *acbA*, *smlA*, *pten*, *abpC* and *modA* had no discernible effect on growth on *Bacillus subtilis* but mutations in *pikA/pikB* and in *cf50-1* caused severe growth defects that were comparable to those seen on *K. pneumoniae*.

Discussion

The results indicate that Collage is capable of prioritizing genes in a reliable manner and identifying genes with various effects on the tested phenotype. This allows the analysis of a broad spectrum of genes in a given biological pathway. Application of the method to this specific question required only a few days of computational work and the validation step required a few more days of work. Considering the low yield of standard genetic screens, it would have taken about a year to identify eight new genes in the bacterial response pathway.

Three out of the five validated bacterial growth genes–*abpC*, *smlA* and *pten*, are involved in actin polymerization and cell motility [33–36]. One explanation for the enrichment of these genes is that the availability of preexisting knockout strains may be enriched with cell motility genes. This is because *D. discoideum* has been used extensively as a model system for chemotaxis, and many genes involved in cell motility have been disrupted and made available to the community. Nonetheless, the importance of actin in the consumption of bacteria may have been previously oversimplified, and the enrichment of these genes could be due to an essential role for actin in bacterial consumption. Proper regulation of actin is required for cell motility, phagocytosis and intracellular trafficking of phagosomes to lysosomes [33–36]. Each of these processes could be important in hunting, consuming and digesting bacteria.

We identified the sugar modifying alpha-glucosidase II enzyme, ModA [37]. Complex sugar modifications are important for biogenesis and intracellular trafficking of proteins. Others have shown that disruption of *modA* results in a lack of anionic N-glycan, which is associated with lysosomal enzymes [38]. While it may not be surprising to identify genes that regulate actin and lysosomes in a direct genetic screen, it is important to see that Collage did so too (S8 Fig).

We also identified one gene, *acbA*, with a less salient relationship to bacterial consumption. Gene *acbA* encodes an Acyl-CoA Binding protein, which is similar to the mammalian diazepam binding inhibitor. Acyl-CoA Binding protein is secreted during *D. discoideum* development and cleaved to form the SDF-2 peptide (Spore Differentiation Factor-2) [39, 40]. The role of Acyl-CoA Binding protein and SDF-2 in growth on bacteria is unclear. It is unlikely to be due to disruption of a general cellular growth pathway, since $acbA^-$ cells grow normally in axenic medium and it is unclear whether the SDF-2 peptide is secreted during growth because the system that produces it is developmentally-regulated. The identification of *acbA* suggests that novel gene functions can be discovered with our gene prioritization method.

The ranking of candidate genes depends on a particular collection of data sets we consider for gene prioritization. Removal of data sets from the data fusion graph (S3 Fig) changes the prioritization. When fewer data sets are considered, the validated genes from our study become ranked lower, below the top 30 (S3 Table). This is an intuitive dependence, less information should result in reduced prioritization accuracy, which we validated by simulations (S9 Fig). For every considered data compendium, Collage achieved a higher area under the ROC (AUC) statistic for known bacterial response genes than for randomly selected genes. However, not surprisingly, the suitability of a data compendium to rank genes depended on the number of data points in the compendium as well as on the usefulness of individual data sets. Our previous computational studies in data fusion with collective matrix factorization bear additional evidence that exclusion of data sets gradually reduces the quality of the predictions [41, 42]. We can attribute our success in identification of genes that participate in Gram-negative response pathways to the proposed approach and the appropriate choice of 14 relevant data sets. In the absence of a much larger set of known genes for this pathway, we cannot claim that this particular selection of data sets is optimal.

Collage builds upon our recently developed data fusion method by collective matrix factorization [23], and extends it with post-processing by latent matrix chaining and gene profiling.

Collective matrix factorization has already provided accurate predictions of gene functions in *Dictyostelium* and yeast [43] and drug toxicity in mouse and human [42], where the accuracy was higher than that of other methods including random forests and approaches based on multiple kernel learning. Another utility of collective matrix factorization was also found in the study of disease interactions [41]. In these studies, collective learning enabled excellent accuracy and effortless integration of a range of very diverse data sets. Collective learning hence provides means for Collage to constitute a useful complement to large-scale ranking of genes in various organisms and to ranking of other objects contained in the fusion graph, such as drugs, diseases and pathways.

Our previous experiments with collective matrix factorization demonstrate that collective matrix factorization applies to diverse range of data sets, learning tasks and organisms. Through latent matrix chaining, Collage adapts collective factorization to prioritization, and thus Collage inherits the general applicability and robustness of collective factorization. Rather than in part extending our previous *in silico* studies we here report on the ability of Collage to make novel and highly accurate predictions.

Materials and Methods

Data sets

A total of 14 data sets and 10 object types were considered for Gram-negative bacterial response gene prioritization. Data sets were organized in a data fusion graph (Fig 2). We used RPKM-normalized RNA-seq transcriptional profiles of 35 abc-transporter mutant strains and wild-type AX4 strain in two biological replicates and at four different time points during development [24] ($\mathbf{R}_{1,8}$), normalized gene expression profiles analyzed by RNA-seq and measured at 4-hour intervals during the 24-hour development of *D. discoideum* in two biological replicates [25] ($\mathbf{R}_{1,7}$), and normalized abundances of gene transcripts in two replicates and four different bacterial growth conditions analyzed with RNA-seq [22] ($\mathbf{R}_{1,9}$). We also included the following publicly available data sets: Phenotype Ontology [29] annotations ($\mathbf{R}_{1,10}$) downloaded from the DictyBase data portal in March 2014, protein-protein interactions from the STRING v.9 database [26] ($\mathbf{\Theta}_1$), membership of *D. discoideum* genes in pathways from the Reactome database [28] ($\mathbf{R}_{1,6}$) downloaded in August 2013, Kyoto Encyclopedia of genes and genomes (KEGG) pathway memberships [27] ($\mathbf{R}_{1,5}$), and annotations of genes in Gene Ontology [30] ($\mathbf{R}_{1,4}$). Additionally, we cross-referenced Reactome and KEGG pathways ($\mathbf{R}_{6,5}$), Gene Ontology terms and Reactome pathways ($\mathbf{R}_{6,4}$), and KEGG orthology groups and Gene Ontology terms ($\mathbf{R}_{5,4}$). Literature data included associations of genes to research articles from PubMed ($\mathbf{R}_{1,2}$) accessed in August 2013 through DictyBase, mapping of research articles to Gene Ontology terms ($\mathbf{R}_{2,4}$) and their Medical Subject Headings (MeSH) ($\mathbf{R}_{2,3}$). As a final step before data analysis, we normalized all relation data matrices such that the Frobenius norm of every row profile was equal to one. S4 Table summarizes the number of objects of each type and the data sets considered in our analysis.

Data fusion by collective matrix factorization

A total of 14 data sets and 10 object types were considered for Gram-negative bacterial response gene prioritization (Fig 2). Data sets are viewed as dyadic relations and are encoded in relation and constraint matrices. A relation matrix $\mathbf{R}_{i,j}$ is a $n_i \times n_j$ real-valued matrix, in which rows correspond to objects of type i, columns to objects of type j and the element $\mathbf{R}_{i,j}(k, l)$ represents the relationship between objects k and l. A constraint matrix $\mathbf{\Theta}_i$ is a $n_i \times n_i$ matrix that relates objects of type i to themselves. It contains pairwise constraints indicating the dissimilarity/similarity between objects. Larger positive elements in $\mathbf{\Theta}_i$ direct data fusion

algorithm to infer a latent model in which the corresponding objects have fewer similar latent profiles (i.e., positive elements in the constraint matrices specify cannot-link constraints). Larger negative elements indicate greater similarity of latent profiles (i.e., negative elements in the constraint matrices specify must-link constraints). Constraint matrices are used for regularization and are not factorized (S4 Fig). Given a collection of relation matrices $\mathcal{R}$ ($\mathbf{R}_{i,j}$ for different choices of i and j) and a collection of constraint matrices $\mathcal{C}$ ($\mathbf{\Theta}_i^{(l)}$ for different choices of i, where l enumerates constraint matrices available for object type i), collective matrix factorization simultaneously decomposes all the relation matrices in $\mathcal{R}$ while regularizing the inferred latent model with the constraints in $\mathcal{C}$. This is accomplished by minimizing our previously proposed loss function [23]:

$$\min_{\mathbf{G}_i \geq 0, \mathbf{S}_{i,j}} \sum_{\mathbf{R}_{i,j} \in \mathcal{R}} \| \mathbf{R}_{i,j} - \mathbf{G}_i \mathbf{S}_{i,j} \mathbf{G}_j^T \|_{\text{Fro}}^2 + \sum_{\mathbf{\Theta}_i \in \mathcal{C}} \sum_{l=1}^{l_i} \text{tr}(\mathbf{G}_i^T \mathbf{\Theta}_i^{(l)} \mathbf{G}_i).$$

The objective function aims at good reconstruction of the observed elements in the data matrices and penalizes violated constraints. The inferred low-dimensional matrix factors $\mathbf{G}_i$ and $\mathbf{S}_{i,j}$ form decompositions of the relation matrices such that $\mathbf{R}_{i,j} \approx \mathbf{G}_i \mathbf{S}_{i,j} \mathbf{G}_j^T$ for all i and j. Here, $\mathbf{G}_i$ is a $n_i \times c_i$ nonnegative latent matrix (a "recipe matrix") containing latent profiles of objects of type i in the rows, $\mathbf{G}_j$ is a $n_j \times c_j$ nonnegative latent matrix with profiles of objects of type j in the rows, and $\mathbf{S}_{i,j}$ is a $c_i \times c_j$ latent matrix (a "backbone matrix") that models interactions between latent components in the (i, j)-th data set. Latent profile of an object of type i is given by its corresponding row vector in $\mathbf{G}_i$ and encodes membership of the object to c_i latent components.

The key principle of data fusion is sharing of latent matrices among decompositions of related matrices. Latent matrix $\mathbf{G}_i$ is utilized for decomposition of any relation matrix that describes objects of type i, that is, $\mathbf{G}_i$ is used in factorizations of matrices $\mathbf{R}_{i,j}$ and $\mathbf{R}_{j,i}$ for any object type j. While latent matrix $\mathbf{G}_i$ is shared, latent matrix $\mathbf{S}_{i,j}$ is specific to the relation $\mathbf{R}_{i,j}$. The inferred latent model thus consists of object type-specific latent matrices ($\mathbf{G}_i$) and latent matrices specific to individual data sets ($\mathbf{S}_{i,j}$).

The algorithm for inference of fused latent models is accompanied by previously reported proofs of correctness and convergence [23]. Briefly, it is an iterative algorithm that starts by randomly initializing latent matrices $\mathbf{G}_i$ and then alternates between updating matrices $\mathbf{G}_i$ and $\mathbf{S}_{i,j}$ until convergence. To ensure robust prioritization, the algorithm was run 20 times with different initializations of latent matrices. The algorithm was run for a maximum of 200 iterations or was terminated early if the total reconstruction error between consecutive iterations changed by less than 0.01. Parameters of the algorithm are factorization ranks, c_i, for every object type i in the data fusion system. Our prioritization of *D. discoideum* genes included 10 types of objects; we have selected latent dimensionality of object types through a single parameter representing the fraction of the original data dimensionality such that $(c_1, c_2, \ldots, c_{10}) = (kn_1, kn_2, \ldots, kn_{10})$. The value of k was obtained by observing kinks in a diagram of total reconstruction error when varying k from 0.05 to 0.5. We selected $k = 0.1$ where a maximum kink was attained. S5 Fig summarizes the procedure and the resulting latent data dimensionality of each object type used in our analysis.

Gene profiling by chaining of latent data matrices

We assembled gene profiles by relying on the latent data matrices inferred by collective matrix factorization. Each gene was characterized through a collection of profiles determined by the topology of data fusion graph. Collage constructed gene profiles by starting at the gene node

and its corresponding recipe matrix ($\mathbf{G}_1$). The method traversed along edges of data fusion graph and multiplied the edge-associated backbone matrices. In the bacterial response gene prioritization study there were 15 chains of latent matrices (Fig 2), and consequently 15 distinct profile matrices containing gene profiles of every considered gene: $\mathbf{G}_1$, $\mathbf{G}_1\mathbf{S}_{1,7}$, $\mathbf{G}_1\mathbf{S}_{1,8}$, $\mathbf{G}_1\mathbf{S}_{1,9}$, $\mathbf{G}_1\mathbf{S}_{1,10}$, $\mathbf{G}_1\mathbf{S}_{1,2}$, $\mathbf{G}_1\mathbf{S}_{1,6}$, $\mathbf{G}_1\mathbf{S}_{1,5}$, $\mathbf{G}_1\mathbf{S}_{1,4}$, $\mathbf{G}_1\mathbf{S}_{1,2}\mathbf{S}_{2,3}$, $\mathbf{G}_1\mathbf{S}_{1,6}\mathbf{S}_{6,5}$, $\mathbf{G}_1\mathbf{S}_{1,6}\mathbf{S}_{6,4}$, $\mathbf{G}_1\mathbf{S}_{1,2}\mathbf{S}_{2,4}$, $\mathbf{G}_1\mathbf{S}_{1,5}\mathbf{S}_{5,4}$ and $\mathbf{G}_1\mathbf{S}_{1,6}\mathbf{S}_{6,5}\mathbf{S}_{5,4}$. It should be noted that latent matrix chains may vary in length and that precise number of chains including a particular backbone matrix is decided by the structure of data fusion graph. For example, in our study, the backbone matrix $\mathbf{S}_{1,6}$ was contained in four chains whereas matrix $\mathbf{S}_{2,3}$ participated in a single chain. Since each resulting profile matrix is determined by a path through object types, adding further away object types increases the weight of intermediate backbone matrices. It therefore can occur that matrices (i.e., data sets), which are present in many chains, have greater influence on prioritization than matrices (i.e., data sets), which appear in fewer chains. However, we would like to note that an intermediate backbone matrix with large latent dimensionality does not necessarily dominate construction of the profile matrix as can be seen from similarity score matrices in S6 Fig. Because Collage operates on matrix chains, it gives a natural approach for incorporating relevance of data sets. Collage assumes that a more relevant object type is the object type that is closer to target type (e.g., genes) in terms of the number of links needed to connect it with the target node. Consequently, in gene prioritization, this means that data sets, which are closely related to genes might have a stronger effect on prioritization than distant non-gene related data sets.

Gene prioritization

The inputs to gene prioritization were candidate genes, seed genes and the set of profile matrices. Collage aims to find genes whose profiles are similar to the profiles of seed genes. The approach estimates the similarities independently for each profile matrix, and then aggregates the resulting scores to obtain the final prioritization. Each row in a profile matrix corresponds to a profile of a gene. Collage assesses similarity between a candidate gene and a seed gene by computing Spearman rank correlation of two respective row vectors. In this study, this procedure yielded a $15 \times |\text{seed genes}|$ similarity score matrix of rank correlations for each candidate gene (S6 Fig). Similarity score matrices are aggregated in a two-step median value computation along score matrix dimensions to produce a single rank value per gene. Collage reports on empirical P-values obtained by randomizing seed set of genes. Randomization of seed genes was repeated 500 times. A nominal P-value of a candidate rank was estimated as $(h+1)/(n+1)$, where n is the number of replicate seed sets that have been simulated and h is the number of these replicates that produced aggregated score greater than or equal to that calculated for the actual seed set.

As a gene profile similarity measure, Collage uses Spearman rank correlation due to the correspondence of rank correlation with assignments of genes to the latent components of inferred matrices. A promising candidate gene should have a latent profile similar to the profile of a seed gene. Given a profile matrix $\mathbf{X}$, candidate gene g and seed gene s, gene g is considered promising if its latent component with the largest membership is the same as that of seed gene s. We formalize this intuition by measuring whether $\arg\max_j \mathbf{X}(g, j) = \arg\max_j \mathbf{X}(s, j)$. The same should hold for the latent component of the second largest, third largest, and all remaining value-ordered gene memberships. Quantitatively, the described procedure corresponds to rank correlations between candidate and seed genes.

The implementation of Collage for bacterial response gene prioritization in *Dictyostelium* is available online (http://github.com/marinkaz/collage). Readers are invited to browse, use and contribute to the software.

Generalization performance of Collage on data subcompendia

To study the sensitivity of gene prioritization to the number of data sets in the data fusion graph, we observed how the rankings of the validated candidate genes changed when the overall prioritization was obtained by fusing different subsets of data sets from our initial collection. In addition to the original model that contained 14 data sets we applied Collage to four independent gene prioritization data scenarios (S3 Table). The scenarios considered seven, four, three and two data sets, where each model was applied to a different subset of entire data collection (S3 Fig). The selection of data sets was in part determined by the data fusion graph. In particular, for data fusion to take place, the associated graph has to be connected such that information can be shared between data matrices. To evaluate the usefulness of Collage to fuse data matrices in a non-gene data space we performed leave-one-out cross-validation. In each validation run, one seed gene was excluded from a set of seed genes and added to test set consisting of *D. discoideum* genes whose knockout strains were available in the Dicty Stock Center. Collage then determined the ranking of this gene for each data scenario separately. From the overall prioritization on a given data compendium, we calculated sensitivity and specificity values of Collage and reported the receiver operating characteristic (ROC) curve and the AUC statistic based on ranks of left-out genes (S9 Fig). As a negative control for prioritization, we applied Collage to randomly selected seed sets of genes using all considered data sets.

Experimental analysis of *Dictyostelium* mutants

D. discoideum strains were obtained from the Dicty Stock Center and grown axenically in HL-5 at 22°C [22]. *K. pneumoniae* was maintained in SM broth at 22°C. To assess the ability of *D. discoideum* to grow on bacteria, *D. discoideum* cells were collected from axenic cultures during logarithmic growth and washed once with Sorensen's buffer [22]. *D. discoideum* cells were serially diluted with bacteria (OD_{600} = 1.0) and spotted onto SM agar plates. The plates were incubated in a humid chamber at 22°C, and images of plates were taken every 24 hours.

Supporting Information

S1 Text. A friendly tutorial to Collage. The tutorial provides a step-by-step explanation of the mathematical and computational concepts considered by Collage.
(PDF)

S1 Fig. A schematic overview of matrix tri-factorization. The figure illustrates the decomposition of the $m \times n$ gene-to-phenotype data matrix **R** into a product of three low-rank latent matrices, **F**, **S** and **G**. The goal of tri-factorization is to approximate the large-scale gene-to-phenotype matrix with a product of much smaller latent matrices such that the approximation is as good as possible. The original $m \times n$ gene-to-phenotype data matrix **R** is compressed by factorization into a much smaller $c_1 \times c_2$ matrix **S** of latent (meta) genes in rows and latent (meta) phenotypes in columns. Matrix **S** is asymmetric and models the interactions between the latent components. To map this compressed representation back to the original domain space, we need two additional matrices, **F** and **G**. The $m \times c_1$ nonnegative matrix **F** maps the space of the meta genes to the space of genes. In each of the m rows, matrix **F** contains the memberships of a respective gene in each of the c_1 latent components (meta genes). Similarly, each column of the $c_2 \times n$ nonnegative matrix **G** contains the memberships of a respective phenotype in each of the c_2 latent components (meta phenotypes).
(PDF)

S2 Fig. Representation of information sources with data matrices. Matrices in Collage describe relationships between objects of two types. Matrix rows correspond to objects of one

type, columns correspond to objects of the other type and matrix elements express the degree of a relationship between the corresponding objects. The figure illustrates matrix representation of six distinct data sets. **(a)** Degrees of protein-protein interactions from the STRING database are represented in a gene-to-gene matrix. **(b)** Membership of genes in pathways are represented in binary matrices, one column for each pathway. Binary matrices are also used to associate **(c)** pathways with gene ontology terms and **(d)** research articles with Medical Subject Headings. **(e)** The structure of Gene Ontology can be represented with a real- valued matrix, whose elements report on distance or semantic similarity between the corresponding ontological terms. **(f)** Levels of gene expression, an experimental data set, are represented by a matrix of stacked gene expression profile vectors.
(PDF)

S3 Fig. Data fusion graphs for the study of model sensitivity to data set selection. Besides the full collection of data sets (data fusion graph in Fig 2), we have considered data collections with a smaller number of data matrices and studied the impact of this reduction on gene prioritization (S3 Table). We ran gene prioritization analyses by considering subsets of **(a)** seven, **(b)** four, **(c)** three and **(d)** two data sets that were included in our original study.
(PDF)

S4 Fig. A schematic overview of penalized matrix tri-factorization. A prominent approach to approximate a matrix with a system of latent matrices is singular value decomposition (SVD). Factorized models inferred by SVD are prone to overfitting, they cannot guarantee conservation of the desired structural properties of the latent matrices, such as nonnegativity, and they are hard to interpret. These shortcomings of SVD and its variants have spurred the development of regularized learning approaches to matrix factorization. Penalized matrix tri-factorization introduces regularization to tri-factorized latent model. In the figure, the input data matrix is accompanied by two constraint matrices that express degrees of similarity between genes (matrix in yellow and orange) or phenotypes (matrix in blue and green). Constraint matrices guide the inference of latent matrices. In our implementation, elements of constraint matrices that have greater negative values represent must-link constraints, i.e., the corresponding genes (or phenotypes) should have more similar latent profiles. Elements with positive values have the opposite effect—they represent cannot-link constraints by penalizing the latent data model if the corresponding genes (or phenotypes) have similar latent profiles. The matrix factorization algorithm balances between good approximation and adherence to the constraints.
(PDF)

S5 Fig. Reconstruction error as a function of factorization rank. Collective matrix factorization requires specification of latent data dimensionality, that is, a factorization rank for each modeled object type. Factorization rank determines the degree of compression of relation matrices: compression is higher with latent matrices of lower dimensionality. The study of bacterial response gene prioritization in *D. discoideum* considered data sets describing relationships between objects of 10 different types. Factorization ranks were set through a single parameter k, where the factorization rank was set to kn_i for each object type i with n_i objects. The value of k was selected by observing the change of the "total reconstruction error" (black line), $\sum_{\mathbf{R}_{i,j}\in\mathcal{R}} \|\mathbf{R}_{i,j} - \hat{\mathbf{R}}_{i,j}\|_{\mathrm{Fro}}$, when varying k between 0.05 and 0.5 (x-axis, "fraction of original data dimensionality"). The reconstruction error was estimated by 50 repetitions of collective matrix factorization, where each repetition was run with a different random initialization of latent matrices. The bars show reconstruction errors of individual data matrices, $\|\mathbf{R}_{i,j} - \hat{\mathbf{R}}_{i,j}\|$

("relation reconstruction error"). See Fig 2 in the main text for description of the data matrices. We selected $k = 0.1$ where the maximum kink was attained. This choice resulted in latent data dimensionality (c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10} = (1287, 342, 280, 308, 9, 9, 5, 28, 5, 50) with a limitation on minimum factorization rank set to 5.
(PDF)

S6 Fig. Visualization of similarity score matrices for candidate genes validated in the wet laboratory. Collage profiles genes through chaining of latent matrices. For a given candidate gene, the profiling procedure yields as many gene profiles (i.e., data vectors corresponding the gene) as there are different chains of latent matrices. Collage then assesses similarity between the candidate gene and a particular seed gene by computing Spearman rank correlation between the respective gene profiles. The figure shows the resulting 15 × 4 (i.e., there were 15 chains and 4 seed genes in our study) similarity score matrix containing rank correlations for each candidate *Dictyostelium* gene that was validated in the wet laboratory.
(PDF)

S7 Fig. Heterogeneity of seed genes. We assessed whether and to what degree the data on seed genes *spc3*, *clkB*, *nip7* and *alyL* that were considered for bacterial response prioritization in *Dictyostelium* vary across individual seed gene. To determine how a given seed gene is different from other seeds, we randomized seed set 400 times and in each randomization used Collage to calculate the similarity score between the given gene and the random set of seed genes (shown in light green). For a given gene we also show its true score as estimated by Collage (e.g., see the vertical line corresponding to the value for *spc3* in top left panel) when only the remaining three seed genes (e.g., *clkB*, *nip7* and *alyL* in top left panel) were considered for scoring. One possible explanation for the substantial amount of variation across seed genes is that these genes were previously identified to be involved in bacterial response pathways using various genetic and genomic methods [22]. They might therefore participate in different aspects of bacterial recognition. Large heterogeneity of seed genes also indicates the difficulty of prioritization task considered here and suggests that consideration of all four seed genes for prioritization is important.
(PDF)

S8 Fig. Homogeneity of candidate genes validated in wet laboratory. We assessed how alike are candidate genes that were validated in wet laboratory on the basis of their latent data representation estimated by Collage. To determine how a given candidate gene is different from other candidates, we randomized set of candidates considered for experimental validation 400 times and in each randomization used Collage to calculate the similarity score between the given gene and the random set of genes (shown in light green). For a given gene we also show its true score as estimated by Collage (e.g., see the vertical line corresponding to the value for *cf50-1* in top left panel) when the remaining seven genes (e.g., *abpC*, *pikA*, *pten*, *modA*, *acbA*, *pikB* and *smlaA* in top left panel) were considered for scoring.
(PDF)

S9 Fig. Generalization performance of Collage. To estimate generalization performance of Collage for bacterial response prioritization in *Dictyostelium*, we performed cross-validation on seed genes in order to obtain sensitivity and specificity of our model. For this task, the leave-one-out cross-validation fitted well. **(Left; a, b)** We applied Collage once for each seed gene using all other seed genes as training genes and the left-out gene as a test gene (positive control). For the negative controls, we considered genes, whose mutants are available in the Dicty Stock Center (727 genes from S2 Table). Shown are the **(a)** ROC curves with the area under the ROC curve statistics, and **(b)** precision-recall curves with the area under the

precision-recall curve statistics based on ranks of left-out genes. The removal of non-gene related data matrices decreased sensitivity and specificity of Collage, suggesting the important ability of Collage to link non-gene related data matrices. The data sources used to construct every performance curve are indicated in S3 Fig. **(Right; a, b) (a)** Rank ROC curves and **(b)** precision-recall curves obtained for the leave-one-out cross-validation performed on eight top ranked candidate genes, which were used for testing Collage and proven to be involved in bacterial response pathways. Notice that higher accuracy of results shown in the right panel relative to results in the left panel was expected as the eight top ranked candidate genes have all been predicted using the same seed set (S8 Fig). We would hence like to warn readers about possible confounding effects present in the experiment whose results are shown in the right panel. In both figures, the control ROC curve (black dashed line) was obtained after prioritization with randomly constructed seed sets and by using all data sources.
(PDF)

S1 Table. Whole-genome prioritization list. Prioritized list of *D. Dictyostelium* genes with the associated empirical P-values and the aggregated prioritization scores as estimated by Collage.
(XLSX)

S2 Table. Gene prioritization list for a subset of genes from the DictyBase available in the Dicty Stock Center. Prioritized list of *D. Dictyostelium* genes with the associated empirical P-values and the aggregated prioritization scores as estimated by Collage. This is the sublist of S1 Table, where only genes that were available in the Dicty Stock Center for direct testing are included.
(XLSX)

S3 Table. The impact of modeling circumstantial data on the overall *D. discoideum* bacterial response gene prioritization. The table lists the top-30 candidate genes obtained by prioritization by data fusion of 14, 7, 4, 3 and 2 data sets from the data fusion graphs in S3 Fig. Genes in bold are the ones selected for the experimental study.
(PDF)

S4 Table. Summary of data sets considered for bacterial response gene prioritization in *D. discoideum*. The notation of the data sets ("Data matrix" column) is the same as in the data fusion graph (Fig 2). All relation data matrices were normalized before data analysis such that the Frobenius norm of every row profile was equal to 1. This type of data normalization was also considered in our previous studies with collective matrix factorization. Preprocessed data sets are provided with the project related code and are available from GitHub repository (http://github.com/marinkaz/collage).
(PDF)

Author Contributions

Conceived and designed the experiments: MŽ GS BZ. Performed the experiments: EAN CD. Analyzed the data: MŽ. Contributed reagents/materials/analysis tools: CD. Wrote the paper: MŽ EAN GS AK BZ.

References

1. Ormrod JE. Human learning. Upper Saddle River, New Jersey, USA: Pearson; 2011.
2. Moreau Y, Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature Reviews Genetics. 2012; 13(8):523–536. doi: 10.1038/nrg3253 PMID: 22751426

3. Franke L, Van Bakel H, Diosdado B, Van Belzen M, Wapenaar M, Wijmenga C. TEAM: a tool for the integration of expression, and linkage and association maps. European Journal of Human Genetics. 2004; 12(8):633–638. doi: 10.1038/sj.ejhg.5201215 PMID: 15114375
4. Sifrim A, Popovic D, Tranchevent LC, Ardeshirdavani A, Sakai R, Konings P, et al. eXtasy: variant prioritization by genomic data fusion. Nature Methods. 2013; 10(11):1083–1084. doi: 10.1038/nmeth.2656 PMID: 24076761
5. Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. Bioinformatics. 2004; 20(16):2626–2635. doi: 10.1093/bioinformatics/bth294 PMID: 15130933
6. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, et al. Gene prioritization through genomic data fusion. Nature Biotechnology. 2006; 24(5):537–544. doi: 10.1038/nbt1203 PMID: 16680138
7. De Bie T, Tranchevent LC, Van Oeffelen LM, Moreau Y. Kernel-based data fusion for gene prioritization. Bioinformatics. 2007; 23(13):i125–i132. doi: 10.1093/bioinformatics/btm187 PMID: 17646288
8. Sun J, Jia P, Fanous AH, Webb BT, Van den Oord EJ, Chen X, et al. A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases–schizophrenia as a case. Bioinformatics. 2009; 25(19):2595–6602. doi: 10.1093/bioinformatics/btp428 PMID: 19602527
9. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009; 37(suppl 2):W305–W311. doi: 10.1093/nar/gkp427 PMID: 19465376
10. Yu S, Tranchevent LC, De Moor B, Moreau Y. Gene prioritization and clustering by multi-view text mining. BMC Bioinformatics. 2010; 11(1):28. doi: 10.1186/1471-2105-11-28 PMID: 20074336
11. Fontaine JF, Priller F, Barbosa-Silva A, Andrade-Navarro MA. Génie: literature-based gene prioritization at multi genomic scale. Nucleic Acids Res. 2011; 39(suppl 2):W455–W461. doi: 10.1093/nar/gkr246 PMID: 21609954
12. Schlicker A, Lengauer T, Albrecht M. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. Bioinformatics. 2010; 26(18):i561–i567. doi: 10.1093/bioinformatics/btq384 PMID: 20823322
13. Sharma A, Chavali S, Tabassum R, Tandon N, Bharadwaj D. Gene prioritization in Type 2 Diabetes using domain interactions and network analysis. BMC Genomics. 2010; 11(1):84. doi: 10.1186/1471-2164-11-84 PMID: 20122255
14. Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet. 2008; 82(4):949–958. doi: 10.1016/j.ajhg.2008.02.013 PMID: 18371930
15. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biology. 2008; 9(Suppl 1):S4. doi: 10.1186/gb-2008-9-s1-s4 PMID: 18613948
16. Mostafavi S, Morris Q. Combining many interaction networks to predict gene function and analyze gene lists. Proteomics. 2012; 12(10):1687–1696. doi: 10.1002/pmic.201100607 PMID: 22589215
17. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nature Methods. 2014; 11:333–337. doi: 10.1038/nmeth.2810 PMID: 24464287
18. Newell P, Henderson R, Mosses D, Ratner D. Sensitivity to Bacillus subtilis: a novel system for selection of heterozygous diploids of Dictyostelium discoideum. Journal of General Microbiology. 1977; 100 (1):207–211. doi: 10.1099/00221287-100-1-207
19. Bozzaro S, Eichinger L. The professional phagocyte Dictyostelium discoideum as a model host for bacterial pathogens. Current Drug Targets. 2011; 12(7):942. doi: 10.2174/138945011795677782 PMID: 21366522
20. Lima WC, Lelong E, Cosson P. What can Dictyostelium bring to the study of Pseudomonas infections? In: Seminars in Cell & Developmental Biology. vol. 22. Elsevier; 2011. p. 77–81.
21. Steinert M. Pathogen–host interactions in Dictyostelium, Legionella, Mycobacterium and other pathogens. In: Seminars in Cell & Developmental Biology. vol. 22. Elsevier; 2011. p. 70–76.
22. Nasser W, Santhanam B, Miranda ER, Parikh A, Juneja K, Rot G, et al. Bacterial discrimination by dictyostelid amoebae reveals the complexity of ancient interspecies interactions. Current Biology. 2013; 23(10):862–872. doi: 10.1016/j.cub.2013.04.034 PMID: 23664307
23. Žitnik M, Zupan B. Data fusion by Matrix Factorization. IEEE Transactions of Pattern Analysis and Machine Intelligence. 2015; 37(1):41–53. doi: 10.1109/TPAMI.2014.2343973

24. Miranda ER, Zhuchenko O, Toplak M, Santhanam B, Zupan B, Kuspa A, et al. ABC transporters in *Dictyostelium discoideum* development. PLoS One. 2013; 8(8):e70040. doi: 10.1371/journal.pone.0070040 PMID: 23967067

25. Parikh A, Miranda ER, Katoh-Kurasawa M, Fuller D, Rot G, Zagar L, et al. Conserved developmental transcriptomes in evolutionarily divergent species. Genome Biology. 2010; 11(3):R35. doi: 10.1186/gb-2010-11-3-r35 PMID: 20236529

26. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013; 41 (D1):D808–D815. doi: 10.1093/nar/gks1094 PMID: 23203871

27. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014; 42(D1):D199–D205. doi: 10.1093/nar/gkt1076 PMID: 24214961

28. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014; 42(D1):D472–D477. doi: 10.1093/nar/gkt1102 PMID: 24243840

29. Fey P, Gaudet P, Curk T, Zupan B, Just EM, Basu S, et al. dictyBase—a Dictyostelium bioinformatics resource update. Nucleic Acids Res. 2009; 37(Suppl 1):D515–D519. doi: 10.1093/nar/gkn844 PMID: 18974179

30. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature Genetics. 2000; 25(1):25–29. doi: 10.1038/75556 PMID: 10802651

31. Chen G, Zhuchenko O, Kuspa A. Immune-like phagocyte activity in the social amoeba. Science. 2007; 317(5838):678–681. doi: 10.1126/science.1143991 PMID: 17673666

32. Zhou K, Takegawa K, Emr SD, Firtel RA. A phosphatidylinositol (PI) kinase gene family in Dictyostelium discoideum: biological roles of putative mammalian p110 and yeast Vps34p PI 3-kinase homologs during growth and development. Mol Cell Biol. 1995; 15(10):5645–5656. PMID: 7565716

33. Gao T, Roisin-Bouffay C, Hatton RD, Tang L, Brock DA, DeShazo T, et al. A cell number-counting factor regulates levels of a novel protein, SslA, as part of a group size regulation mechanism in Dictyostelium. Eukaryot Cell. 2007; 6(9):1538–1551. doi: 10.1128/EC.00169-07 PMID: 17660362

34. Dormann D, Weijer G, Dowler S, Weijer CJ. In vivo analysis of 3-phosphoinositide dynamics during Dictyostelium phagocytosis and chemotaxis. Journal of Cell Science. 2004; 117(26):6497–6509. doi: 10.1242/jcs.01579 PMID: 15572406

35. Cox D, Wessels D, Soll D, Hartwig J, Condeelis J. Re-expression of ABP-120 rescues cytoskeletal, motility, and phagocytosis defects of ABP-120-Dictyostelium mutants. Molecular Biology of the Cell. 1996; 7(5):803–823. doi: 10.1091/mbc.7.5.803 PMID: 8744952

36. Brock DA, Hatton RD, Giurgiutiu DV, Scott B, Ammann R, Gomer RH. The different components of a multisubunit cell number-counting factor have both unique and overlapping functions. Development. 2002; 129(15):3657–3668. PMID: 12117815

37. Ebert DL DR Bush JM, JA C. Biogenesis of lysosomal enzymes in the alpha-glucosidase II-deficient modA mutant of Dictyostelium discoideum: retention of alpha-1,3-linked glucose on N-linked oligosaccharides delays intracellular transport but does not alter sorting of alpha-mannosidase or beta-glucosidase. Arch Biochem Biophys. 1989; 273:479–490.

38. Hykollari A, Dragosits M, Rendić D, Wilson IB, Paschinger K. N-glycomic profiling of a glucosidase II mutant of Dictyostelium discoideum by "off-line" liquid chromatography and mass spectrometry. Electrophoresis. 2014; 35(15):2116–2129. doi: 10.1002/elps.201300612 PMID: 24574058

39. Cabral M, Anjard C, Loomis WF, Kuspa A. Genetic evidence that the acyl coenzyme A binding protein AcbA and the serine protease/ABC transporter TagA function together in Dictyostelium discoideum cell differentiation. Eukaryot Cell. 2006; 5(12):2024–2032. doi: 10.1128/EC.00287-05 PMID: 17056744

40. Cabral M, Anjard C, Malhotra V, Loomis WF, Kuspa A. Unconventional secretion of AcbA in Dictyostelium discoideum through a vesicular intermediate. Eukaryot Cell. 2010; 9(7):1009–1017. doi: 10.1128/EC.00337-09 PMID: 20472692

41. Žitnik M, Janjić V, Larminie C, Zupan B, Pržulj N. Discovering disease-disease associations by fusing systems-level molecular data. Scientific Reports. 2013; 3. doi: 10.1038/srep03202 PMID: 24232732

42. Žitnik M, Zupan B. Matrix factorization-based data fusion for drug-induced liver injury prediction. Systems Biomedicine. 2014; 2(1):16–22. doi: 10.4161/sysb.29072

43. Žitnik M, Zupan B. Matrix factorization-based data fusion for gene function prediction in baker's yeast and slime mold. In: Pacific Symposium on Biocomputing. vol. 19; 2014. p. 400–412.