Gene set enrichment analysis using RNA-seq-based condition-specific metabolic networks

Van Du T. Tran[1], Sébastien Moretti[1], Alix T. Coste[2], Sara Amorim-Vaz[2], Dominique Sanglard[2], Marco Pagni[1*]

[1] Vital-IT Group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[2] Institute of Microbiology, University Hospital Lausanne and University Hospital Center, Lausanne, Switzerland

Running title: Gene set enrichment on metabolic network

Keywords: genome-scale metabolic network, RNA-seq, gene ontology, gene set enrichment, *Yarrowia lipolytica*.

* Corresponding author:

Marco Pagni, Vital-IT group, SIB Swiss Institute of Bioinformatics, Quartier Sorge, bâtiment Génopode, 1015 Lausanne, Switzerland. Email: marco.pagni@sib.swiss

Abstract

Genome-scale metabolic networks and transcriptomic data represent complementary sources of knowledge about an organism metabolism. We propose here a method for gene set enrichment analysis, metaboGSE, to interpret transcriptomic data with the help of a genome-scale metabolic network. The method relies on the creation of series of condition-specific metabolic submodels using expression data. The submodels are assessed by a newly defined fitness function, which then serve to contextualize any gene sets of interest. The functional discrepancy between multiple conditions is then evaluated with a statistical test. The method was validated using publicly available data for *Yarrowia lipolytica* and produced more informative and detailed results with respect to metabolism comparing to GSEA or topGO.

Introduction

The advent of high throughput sequencing techniques, especially RNA sequencing has greatly facilitated the experimental investigation of an organism transcriptome in different physiological conditions. RNA-seq data consists in reads mapped onto an annotated genome, which permits to quantify the abundances of transcripts predicted from all genes. These values can be used as a proxy to quantify gene expression, or may provide hints about protein abundances for protein-coding genes. Differential expression between two conditions or co-expression profiles across many conditions are currently the fundamental statistical approaches to analyze RNA-seq data. However, to obtain a biological interpretation from these analyses, genes also need to be carefully annotated with prior biological knowledge. For example, the Gene Ontology (GO) annotates genomes by grouping genes into sets, each one having a common GO term corresponding to a particular function or sub-cellular location. GO terms are hierarchically arranged in a directed acyclic graph (DAG) whose structure can be taken into account via a computational method such as topGO (Alexa & Rahnenführer, 2016). Another popular method, GSEA (Subramanian *et al*, 2005) alternatively attempts to identify enriched GO terms by considering particular ordering of the genes established from the differential expression values.

Annotated genomes can also serve as a basis to metabolic reconstruction, hereafter referred to as genome-scale metabolic network (GSMN). GSMNs have been widely used to study and model metabolism in living organisms (Feist & Palsson, 2008; McCloskey *et al*, 2013; Oberhardt *et al*, 2009). They are complex biological networks that include thousands of interconnected nodes of four main types: metabolites; biochemical and transport reactions; enzymes and transporters; and genes. The basic building block of a GSMN is the gene-protein-reaction-metabolite tetrad that relates specific genes to their protein products and describes their functions down to the chemical level. A GSMN can be turned into predictive model, after defining (i) the model environment, *i.e.* the growth medium and the excreted metabolites through *external reactions* and (ii) an artificial *growth* reaction which represents how the biomass is assembled at the expense of consumed metabolites. Flux Balance Analysis (FBA) (Varma & Palsson, 1994) can then be used to establish if biomass can be produced in a specific growth medium, *i.e.* to establish the *viability* of the model. A viable GSMN can therefore be used to predict essentiality of genes for example, by simulating the effect on viability of gene knockout experiments, which can be validated against experimental data (Imam *et al*, 2015; O'Brien *et al*, 2015; Simeonidis & Price, 2015).

Several research groups have attempted to integrate RNA-seq data and GSMN. The different published methods have been recently reviewed (Machado & Herrgård, 2014; Kim & Lun, 2014; Vivek-Ananth & Samal, 2016) and can be classified in two broad categories: those aimed at improving the prediction of FBA flux distribution in the network and those aimed at extracting context-specific sub-networks. Regarding the prediction of flux distributions, the methods include Akesson *et al* (Akesson *et al*, 2004), Moxley *et al* (Moxley *et al*, 2009), E-Flux (Colijn *et al*, 2009; Brandes *et al*, 2012), E-Flux2 (Kim *et al*, 2016), PROM (Chandrasekaran & Price, 2010), MADE (Jensen & Papin, 2011), tFBA (van Berlo *et al*, 2011), Lee *et al* (Lee *et al*, 2012), Fang *et al* (Fang *et al*, 2012), RELATCH (Kim & Reed, 2012), TEAM (Collins *et al*, 2012), GX-FBA (Navid & Almaas, 2012), FCGs (Kim *et al*, 2013b), iMAT (Zur *et al*, 2010), GIMME (Becker & Palsson, 2008), GIM3E (Schmidt *et al*, 2013, 3), and EXAMO (Rossell *et al*, 2013). They were reviewed and benchmarked by Machado and Herrgård (Machado & Herrgård, 2014), who concluded with two ambiguous observations: absolute and relative gene expression seem to perform similarly; using gene expression as additional FBA constraints is not clearly superior to what obtained with *parsimonious FBA*, which does not depend on transcriptomic data at all. The methods for construction of context-specific models include iMAT, GIMME, GIM3E, EXAMO, INIT (Agren *et al*, 2012), AdaM (Töpfer *et al*, 2012), mCADRE (Wang *et al*, 2012), GAM (Sergushichev *et al*, 2016), and RegrEx (Robaina Estévez & Nikoloski, 2015). They consist in removing genes which products are supposed to be absent based on the context-specific gene expression. They differ in the way they dealt with the following recurrent problems: how to choose the genes/reactions to remove given the RNA-seq data; how many genes to remove; how to deal with unviable submodels caused by the removal of an essential reaction. None of them seem to be definitely addressed. Moreover, little biological interpretation from contrasting such context-specific sub-networks have arisen, apart from predicting and validating the flux distribution of some particular metabolites of interest (Machado & Herrgård, 2014).

Here we describe a novel method that incorporates both gene expression data and GSMN structure for gene set enrichment analysis. Gene expression data is not used as direct input for gene set enrichment analysis but rather as a means to create a *series* of metabolic sub-networks contextualized to the gene set under study and which are ranked using a newly defined *fitness* function. Our method then considers how a GO term enrichment score evolves with the fitness of a given metabolic network. By utilizing the structure of the GSMN, our method appears to provide a more sensitive way to extract functional categories from gene expression data than lists of differentially expressed genes. The *metaboGSE* R package is available from CRAN.

To validate the method we used existing experimental data and a metabolic model for *Yarrowia lipolytica*, a yeast which is widely used in industrial microbiology for the production of fatty acids (Ledesma-Amaro & Nicaud, 2016). The metabolic model iMK735 of *Y. lipolytica* (Kavšček *et al*, 2015) was exploited, as it can simulate the growth and production of lipids when oxygen consumption is limiting. Gene expression data was taken from the study of Maguire *et al.* on the role of the sterol-regulatory element binding protein Sre1 and the transcription factor Upc2 in sterol metabolism in hypoxic and normoxic conditions (Maguire *et al*, 2014). This study showed that the control of the sterol pathway has been taken over by Upc2 in *Y. lipolytica* and the *Saccharomycotina* group, unlike most eukaryotes, including the majority of fungi.

Results

1. Datasets

The investigated dataset is available from the R package *metaboGSE*.

a. Genome, proteome, GO annotations and genome-scale metabolic network

The genome GCA_000002525.2_ASM252v1 of *Y. lipolytica* was used throughout this study. Its proteome and the list of associated GO terms were retrieved from UniProt (UniProt Consortium, 2015). The *Y. lipolytica* iMK735 model (Kavšček *et al*, 2015) with a production of 40% lipid content in the biomass was studied. The genome, the proteome, the GO annotations, and the model were integrated within the framework of MetaNetX (Moretti *et al*, 2016). The external reactions of the model were adapted to approximately simulate growth in YPD medium in hypoxic (anaerobic) or normoxic (aerobic) environments (Maguire *et al*, 2014) by modifying the oxygen supply. The model was cleaned by removing dead-end metabolites, which were only either produced or consumed. *Blocked* reactions as given by flux variability analysis (FVA) (Mahadevan & Schilling, 2003) were also removed, as explained below. The finally investigated model contained 946 reactions, 504 genes, and 685 metabolites after cleaning, and is referred to as the *comprehensive* model in our analysis.

The GO annotation of *Y. lipolytica* consisted of 4629 GO terms and 4747 annotated genes (out of 6453). An initial enrichment was performed via topGO with *weight01* algorithm and *fisher* statistic on the set of genes from the model iMK735 *versus* the complete genome (gene universe). This yielded 142 GO terms (denoted *GO142* and listed in Appendix Table S1), which belonged to the biological process category and were associated with at least 3 and at most 50 genes in the iMK735 model (enriched with p-value < 0.1). The comparison of our algorithm, topGO, GSEA, and Maguire *et al*. was performed on this dataset, that we assume to contain most relevant information documented by iMK735.

b. RNA-seq data

We investigated the RNA-seq data obtained from Maguire *et al* (NCBI: PRJNA205557) with 22 samples from seven conditions of normoxic and hypoxic growth with *sre1* Δ, *upc2* Δ, *sre1* Δ/*upc2* Δ mutants and the wild type strain (Maguire *et al*, 2014) (see Table 1 and standard RNA-seq analysis in Materials and Methods). The double mutant *sre1* Δ/*upc2* Δ in hypoxic condition is absent from the experimental design as it failed to grow. The two mutated genes *SRE1* and *UPC2* are absent from the iMK735 model, which is expected, as they are transcription factors.

2. Measure of metabolic model fitness

Growth capacity, *i.e.* viability is crucial for the usability of a metabolic model. Removing a few reactions from a viable network is usually sufficient to render it unviable. Here we propose a measure to assess how close an unviable model is from a viable one, which will be called the *fitness* of the model.

a. FBA, FVA and blocked reactions

These classic algorithms are summarized here as used in this paper. Let $\mathcal{M}$ be a GSMN with n reactions, m metabolites. $\mathbf{S}$ denotes the stoichiometric matrix of dimension $m \times n$, which contains the stoichiometric coefficients of metabolites in each reaction, with positive coefficients for products and negative coefficients for substrates. Let $\boldsymbol{v}$ be a vector representing the fluxes on the n reactions, which are constrained by lower and upper bounds, given by the vectors **lb** and **ub**, respectively. Assuming that the flux distribution exists in a quasi-steady state, *i.e.* the rate of production of each metabolite is equal to its rate of consumption. The balance in the system is expressed as

$$\mathbf{S} \cdot \boldsymbol{v} = \mathbf{0}.$$

FBA uses linear programming (LP) to optimize an objective function $\mathbf{c}^{\mathrm{T}} \cdot \boldsymbol{v}$ where $\mathbf{c}$ is a vector of weights. Maximizing the biomass production v_{Biomass} is of interest here and can be achieved by formulating the particular $\mathbf{c}_{\mathrm{Biomass}}^{\mathrm{T}}$ vector with a weight equal to 1 for the growth reaction and weights equal to 0 for all the other reactions. In this context, the optimization problem addressed by FBA can be stated as

$$\text{find} \quad v_{\mathrm{Biomass}}^{*} = \max_{\boldsymbol{v}}\left(\mathbf{c}_{\mathrm{Biomass}}^{\mathrm{T}} \cdot \boldsymbol{v}\right) \quad \text{subject to} \quad \mathbf{S} \cdot \boldsymbol{v} = \mathbf{0} \text{ and } \mathbf{lb} \leq \boldsymbol{v} \leq \mathbf{ub}$$

and the model is said to be *viable* if $v_{\mathrm{Biomass}}^{*} > 0$.

FVA can be used to detect the reactions that are *blocked* in the model, *i.e.* they cannot carry a flux and can be removed from the model without affecting its predictive properties. It consists in solving the following LP problem:

$$\begin{aligned} &\text{find } \min v_i \text{ and } \max v_i \text{ for } i = 1..n \\ &\text{subject to} \quad \mathbf{S} \cdot \boldsymbol{v} = \mathbf{0} \\ &\qquad\qquad\quad \mathbf{lb} \leq \boldsymbol{v} \leq \mathbf{ub} \\ &\qquad\qquad\quad 0 < \mathbf{c}_{\mathrm{Biomass}}^{\mathrm{T}} \cdot \boldsymbol{v} \leq v_{\mathrm{Biomass}}^{*} \end{aligned}$$

The ith reaction is considered to be blocked if $\min v_i = \max v_i = 0$ (Henry *et al*, 2007).

Submodel construction by removing one or several genes refers to the process of (i) removing an initial set of genes (input) and their associated reactions, (ii) determining the blocked reactions, (iii) removing the latter with their associated genes, if any. A similar procedure *deleteModelGenes* was introduced in the COBRA Toolbox (Becker *et al*, 2007). Below, we are referring to a particular submodel with the number of initially removed genes, but all analysis results that are presented have been obtained after gene removal propagation through the blocked reactions.

b. Principle of growth rescue

In this section, we introduce a procedure to restore the viability of an unviable metabolic network by the introduction of a minimal set of artificial reactions. It consists first in modifying the input network around the growth equation as depicted in Fig 1. An artificial metabolite is created for each of those present in the growth reaction. Each artificial metabolite x' can be produced or consumed via a *rescue* external reaction denoted r_x. Each artificial metabolite x' is linked to the original metabolite x through a directed *help* reaction denoted h_x. The purpose of this *help* reaction

is to avoid artificially supplying x' as a side effect of rescuing x. No restriction is applied on the fluxes of the rescue and help reactions apart from the directions. In addition, the same network modifications were applied for non-growth associated maintenance reactions since they do not allow for a zero flux.

Let $\mathcal{M}'$ be the expanded version of $\mathcal{M}$ with the modified growth equation and the full set of rescue and help reactions, $\mathbf{S}'$ is the corresponding stoichiometric matrix. The flux distribution $\boldsymbol{v}'$ is constraint by the bound vectors $\mathbf{lb}'$ and $\mathbf{ub}'$ that account for the directionality of newly added reactions and constraint the growth equation at a fixed rate arbitrarily chosen as 20% the original model growth objective, *i.e.* $\mathrm{lb}'_{\mathrm{Biomass}} = \mathrm{ub}'_{\mathrm{Biomass}} = \frac{1}{5} v^*_{\mathrm{Biomass}}$.

Let $\mathbf{c}_{\mathrm{Rescue}}$ be a vector of weight that is equal to 0 for all reactions but the rescue reactions. It is given a value of $1/B_x$ where B_x is the stoichiometric coefficient for the metabolite x in the original growth equation. The rescue procedure can be stated as the following LP problem:

$$\text{find } \boldsymbol{v}'^* = \mathrm{argmin}_{\boldsymbol{v}'}\left(\mathbf{c}^{\mathrm{T}}_{\mathrm{Rescue}} \cdot \boldsymbol{v}'\right)$$

$$\text{subject to } \mathbf{S}' \cdot \boldsymbol{v}' = \mathbf{0}$$

$$\mathbf{lb}' \leq \boldsymbol{v}' \leq \mathbf{ub}'.$$

Only rescue reactions for metabolite x with non-zero flux $\left|v'^*_{r_x}\right| > 0$ are required to restore the model viability. The fraction of metabolites in the growth reaction that need to be rescued could have constituted a possible measure of the model fitness. We investigated this by constructing submodels through the random withdrawal of variable numbers of genes (from 1 to the total number of genes). Random draws were performed $N = 50$ times. The simulation shows the crucial property that the more genes removed, the more damaged the model (Fig 2A). Fig 2B presents the fraction of draws where each individual metabolite is rescued, clearly indicating different behaviors among metabolites in the growth equations. Furthermore, our interest here is focused on not-too-damaged submodels, say, submodels obtained after removal of up to 20% of genes. This suggests that the different metabolites should not be treated equivalently.

c. Weighting schema for model fitness

The idea here is to introduce a weighting scheme to account for the variable importance of metabolites in the growth equation and possible dependencies among them (Fig 2B). Model fitness is defined as

$$F(\mathcal{M}') = 1 - \frac{\sum_{r_x} w_{r_x} \mathbb{I}\left(\left|v'^*_{r_x}\right| > 0\right)}{\sum_{r_x} w_{r_x}},$$

where $\mathbb{I}(e) = 1$ if e is true, and 0 otherwise. The weights w_{r_x} are computed as follows. For every gene in the model, a single gene knockout is simulated and the rescue procedure is performed to determine which metabolites in the growth equation are affected. Hence, every rescue reaction r_x is associated with a binary vector which describes whether the reaction is necessary or not for each of the gene knockouts. These binary vectors are used to compute Euclidean distances between rescue reactions, which are used in hierarchical clustering with average linkage, and the Gerstein method [41] is then applied to the resulting tree as a means to assign a weight to each

rescue reaction. Those rescue reactions with similar binary vectors share weights, while those with unique profile receive a larger weight. The weighting schema has two main effects: i) it reduces the importance of metabolites that are weakly affected by slightly damaging the model, and ii) assigns similar importance to metabolites that appear on the same pathway, as shown in Fig 2C. For instance, H_2O has the smallest weight, phosphatidylcholine and phosphatidylethanolamine in the same pathway share weights. With the simulation of random draws above, the distribution of obtained submodel fitness scores was less dispersed than that of fraction of rescued metabolites (see Fig 2A and 3).

3. Optimal ranking of genes for removal

A condition-specific metabolic submodel can be constructed by removing a certain number of genes or reactions from the comprehensive network. Its health can be evaluated via the fitness metric proposed in the previous section. The question then arises as to how to select genes for elimination in a particular condition, with the aim of removing more genes while less disturbing the network or more preserving its high fitness. In this section, we investigate a number of strategies to rank genes for removal according to their expression level.

We tested the following transformations of the raw expression data to rank the genes:

expr	average $\log_2$-counts of genes per condition
pkmExpr	expression in RPKM (see Material and Methods)
relExpr1 = **expr**/<**expr**>	relative expression
relExpr2 = (**expr**^{2}/<**expr**>)$^{1/2}$	another relative expression
relExpr3 = (**expr**^{3}/<**expr**>)$^{1/3}$	yet another relative expression
zExpr = (**expr** - <**expr**>)/sd(**expr**)	*z*-score
revExpr = 1/(1+**expr**)	reversed expression (control),

where <**expr**> and sd(**expr**) denote the average and standard deviation across conditions, respectively. Let ρ be a particular ranking of the genes based on their expression level. Submodels are constructed by removing genes in the order given by ρ. The produced fitness scores decrease when more genes are removed and the decreasing trend depends on ρ. Fig 3 presents the fitness of the submodels obtained by successively removing genes from the comprehensive model iMK735 for the UH condition (*upc2*Δ in hypoxic condition) and for the seven rankings described above.

We define a *significance level* $\boldsymbol{Z}_\rho$ for the ranking ρ, which corresponds to the percentage of random draws (as in Section 2) in which submodel fitness is higher than that of submodels created by ρ-based removal of the same number of removed genes and which is weighted by the average random fitness.

$\boldsymbol{Z}_\rho$ is computed as follows:

$$\boldsymbol{Z}_\rho = 100\times \sum_i \left(\frac{\langle F\left(\mathbf{M}_i^{random}\right)\rangle}{\sum_i \langle F\left(\mathbf{M}_i^{random}\right)\rangle} \sum_{N \text{ draws}} \frac{1}{N} \mathbb{I}\left(F\left(\mathbf{M}_i^{random}\right) \geq F\left(\mathbf{M}_i^{\rho}\right)\right)\right),$$

where $\mathbf{M}_i^{\rho}$ denotes the rescued submodel of the comprehensive network $\mathbf{M}$ by removing the first i genes following the ranking ρ, and $\mathbf{M}_i^{random}$ the rescued submodel by randomly removing i genes, $\langle F(\ldots)\rangle$ the average fitness on N draws. The lower the score, the better the ranking.

The gene removal following absolute expression ranking **expr** produced the significance level $\boldsymbol{Z}_{\mathbf{expr}}$ of 2.69% on Fig 3. Table 2 shows that the gene ranking on **expr** outperforms the others with $\boldsymbol{Z}_{\mathbf{expr}}$ ranging between 2.49 and 7.91% for all conditions (see also Appendix Fig S1). **expr** and **relExpr*i*** (*i* = 1,2,3) shows some noteworthy distinction in their scores, especially between **expr** and **relExpr1**. The reverse ranking (**revExpr**) was introduced as a control. Its high $\boldsymbol{Z}_{\mathbf{revExpr}}$ value indicates that the reverse ranking is clearly worse than a random gene removal. The **expr** ranking was then used in the rest of this study.

4. Condition-specific submodel construction and comparison of multiple conditions

Now we apply the optimal gene removal strategy to construct submodels for the seven conditions studied by Maguire *et al*. The gene expression profiling and hence the **expr** ranking are specific to each condition. As a result, the submodels constructed based on **expr** evolve differently across conditions, and so their fitness also evolves differently (Fig EV1A,B). A *stable* region of fitness higher than 0.9 can be observed, which reflects genes of very low expression. This suggests that we can safely remove a portion of lowly expressed genes without much affecting model properties. This region also shows a relatively stable percentage of essential reactions across conditions, despite a high degree of variation in the proportion of essential genes (Fig EV1C,D) indicating that the submodels are structurally distinct across conditions. The UpSet plot (Lex *et al*, 2014) in Fig EV2 describes the distribution of removed genes before and after propagation when removing 12.5% of the whole model gene set (63 genes) and confirms the difference between conditions.

5. MetaboGSE: contrasting gene set enrichment in condition-specific submodels

We introduce here the metaboGSE method, which identifies gene sets that are differentially enriched across conditions using submodels derived from a comprehensive GSMN. The procedure consists in three main steps depicted on Fig 4:

i. Construction of successive submodels, taking as input a GSMN and a matrix of absolute gene expression **expr** (see Fig EV1A) to rank the genes and produce a matrix of fitness *F* (see Fig EV1B), as described above.

ii. Computation of the fraction *f* of genes associated with a given gene set *g* remaining in each submodel and for each condition. The *enrichment score* of *g* is defined as this fraction, weighted by the fraction of remaining genes and fitness of the corresponding submodel,

$$E_i^g = \left(1 - \frac{i}{k}\right) \cdot F\left(\mathbf{M}_i^{\mathbf{expr}}\right) \cdot f(\mathbf{M}_i^{\mathbf{expr}}, g),$$

where $\mathbf{M}_i^{\mathbf{expr}}$ denotes the rescued submodel of the comprehensive network **M** by removing the first *i* genes and *k* the total number of genes in **M**. A value of fitness may correspond to several enrichment scores. To obtain distributions of enrichment scores as a function of submodel fitness for each condition, a linear interpolation with mean ties is subsequently performed at *l* equally distributed values of fitness, where *l* is the number of unique values of strictly positive fitness obtained in all the conditions. This produces a matrix of interpolated enrichment scores that represents the evolution of *g* when reducing model fitness, for all the conditions

iii. The Kruskal-Wallis statistic was applied on the enrichment score distributions to produce a *p*-value that indicates whether the enrichment evolution of *g* from at least one condition stochastically dominates those from the others. The *p*-values were subsequently adjusted by Benjamini-Hochberg (BH) correction on all the studied gene sets (Benjamini & Hochberg, 1995).

This procedure is illustrated in Fig 5 with GO:0009086 – methionine biosynthetic process. The fraction *f* of genes associated with GO:0009086 decreases rapidly in all conditions (Fig 5A), while the enrichment scores as a function of submodel fitness clearly separate normoxia and hypoxia (Fig 5B). Such discrepancy between conditions is amplified in Fig 5B with regards to that in Fig 5A, and can be captured via the Kruskal-Wallis *p*-value. The normoxia-hypoxia separation can also be found in the hierarchical clustering based on the expression profile of all the 14 genes (Fig 5C), which are not restricted to differentially expressed genes (Fig 5D).

6. Biological interpretation

To validate the biological findings produced by our approach, we investigate in this section the gene sets defined by *GO142*. Maguire *et al*. (Maguire *et al*, 2014) studied the role of Sre1 and Upc2 in regulating sterol metabolism in hypoxic and normoxic conditions in *Y. lipolytica* by performing GO term enrichment analysis of differentially expressed genes using DAVID. Among the 116 biological-process GO terms they reported, only eight are retrieved in *GO142*, but none has a reported raw *p*-value < 0.05. Here we are comparing our results with those of topGO (Alexa & Rahnenführer, 2016), using two algorithms *classic* and *weight01*, and GSEA (Subramanian *et al*, 2005) (see details in Materials and Methods).

Hypoxia-normoxia contrast

Thirty-six GO terms were found to be significantly enriched according to metaboGSE, with BH-adjusted *p*-values of < 0.05 (see Table EV1). Enrichment score curves (Fig S2) indicate that the degree of oxygen limitation is the most likely explanation for the enrichment of 22 of the 36 GO terms, which is expected for this dataset – where hypoxia was the only environmental variable. In addition to methionine biosynthetic process (GO:0009086), three more terms related to sulfur metabolism are found: GO:0019344 (cysteine biosynthetic process, GO:0000103 (sulfate assimilation) and GO:0046938 (phytochelatin biosynthetic process). Interestingly, several terms related to urea cycle are enriched and include GO:0019627 (urea metabolic process), GO:0006591 (ornithine metabolic process), and GO:0006526 (arginine biosynthetic process). Moreover, GO:0009068 (aspartate family amino acid catabolic process) and those concerned in pyrimidine metabolism (GO:0006207, GO:0009221, GO:0009130) might also be related to urea cycle.

Specificity of GO term enrichment

The GO terms found to be significantly enriched by metaboGSE, topGO, and GSEA are summarized in Fig EV3 and Table EV1. Of the 36 GO terms found to be significantly enriched by metaboGSE, 20 are child terms of those found by the other methods, 2 are themselves linked by parent-child relations, and 14 are unlinked by the DAG structure of the GO. Terms found to be significantly enriched by metaboGSE were never found as parents of the terms found only by topGO and/or GSEA. All terms identified via metaboGSE and most of terms identified via topGO or GSEA were either unrelated or children of those reported in Maguire *et al*. Overall, metaboGSE seems to produce more specific terms than the other methods.

An example is shown in Fig 6 for the histidine and purine biosynthesis pathways, which have been described to be co-regulated in *Saccharomyces cerevisiae* (Rébora *et al*, 2005). Whereas GO:0000105 (histidine biosynthetic process) was enriched by both metaboGSE and GSEA, GO:0006189 ('de novo' IMP biosynthesis) was enriched only via metaboGSE. Interestingly, four GO terms that are parents of GO:0006189 were also found with GSEA. This suggests that a similar cross regulation between the two pathways occur in *Y. lipolytica*. Both metaboGSE and GSEA were able to capture it, though metaboGSE provided more informative results. Regarding the term missed by metaboGSE, three were enriched by all the other methods, including GO:0006637 (acyl-coA metabolic process), GO:0035428 (hexose transmembrane transport), GO:0046323 (glucose import), but these are not very informative term.

Sterol biosynthesis

Ergosterol biosynthetic process (GO:0006696) was enriched by all methods. Along with GSEA, we also discovered GO:0045337 (farnesyl diphosphate biosynthetic process), which is part of the ergosterol synthesis pathway (see Appendix Fig S3). This reflects the experimental design investigated in Maguire *et al*., i.e. the regulation of sterol metabolism. In Fig EV4A,B, the sudden drop from 100% to 0% of ergosterol-associated genes indicates that either all these genes or none of them exist in each submodel. This is the consequence of the linearity of the pathway of ergosterol biosynthesis, which is not perceptible in the expression profile of the genes (Fig EV4C). This co-participation of ergosterol-associated genes can only be captured via metaboGSE with all the 19 genes in iMK735 associated with ergosterol biosynthesis, whereas GSEA and topGO captured 16 and 4 genes, respectively. Besides, Maguire *et al*. only detected sterol biosynthetic process (GO:0016126) with two genes from the model (see Fig EV4D). Fig EV4B shows the discrepancy between hypoxia and normoxia in enrichment evolution of ergosterol-associated genes. Moreover, the grouping of *upc2Δ* and wildtype in both hypoxic and normoxic conditions as well as *sre1Δ/upc2Δ* and *sre1Δ* in normoxia and discrepancy triggered by Sre1 deletion reveal the role of Sre1, but not Upc2, in ergosterol regulation in the framework of iMK735. The displacement of Sre1 by Upc2 in sterol regulation claimed in Maguire *et al*. is not confirmed by our analysis, but this particular question might be somehow out of scope as only two genes (YALI0E32065g and YALI0F08701g) associated to sterol biosynthetic process reported in Maguire *et al*. belong to iMK735.

Terms not contrasted by hypoxia-normoxia

metaboGSE detected GO:0001676 (long-chain fatty acid), GO:0006635 (fatty acid beta-oxidation), GO:0015940 (pantothenate biosynthetic process), and GO:0046938, which are the most specific terms in the fatty acid metabolism compared to those resulting from other methods (see Appendix Fig S4). This reflects the nature of the iMK735, which simulates lipid production. A distinction between WH and the remaining conditions can be observed for the two terms on fatty acid, while SH is more distinguished from the others for pantothenate biosynthesis. Additionally, tryptophan biosynthetic process and chorismate metabolic process also revealed a discrepancy between WH and the others. These findings might suggest a role of Sre1 and/or Upc2 in low-oxygen growth condition for the corresponding biological processes.

Discussion

We present here a method for gene set enrichment analysis that utilizes a GSMN as an additional source of information and that focuses on genes expressed at low levels. Our central working hypothesis is that the correlation between gene expression levels and fluxes on related reactions is very poor in general, but the lowly expressed genes are plausibly associated with zeroed fluxes. This method is complementary to classical methods such as topGO and GSEA that focus on differential expression of sufficiently expressed genes. The introduction of a GSMN restricts the list of investigated genes to those related to metabolism and thus the list of gene sets that can be discovered. Clearly, the quality and scope of the investigated GSMN is likely to affect the method outcome, although this has not been investigated here. The GSMN and the set of RNA-seq data both need to be of high quality and mutually coherent with respect to experimental designs. Our method seems capable of producing more informative GO terms that are located lower in the Gene Ontology than those returned by GSEA and topGO for example. This might be related to the fact that metaboGSE can increase the size of investigated gene sets by taking into account structural constraints brought by the GSMN, as for example the linearity of the ergosterol biosynthesis pathway. The genes affecting the discrepancy between conditions, which are not necessarily differentially expressed, for each enriched gene set can be further investigated. metaboGSE seems to produce biologically meaningful results to the extent one can interpret them.

Our method does not aim at producing a particular condition-specific submodel, but rather integrates on a range of them, thus avoiding the choice of a particular number of genes to remove. A GSMN is a drastic simplification of our understanding and knowledge of biochemistry that neglects most kinetic aspects in its representation of metabolism. On the modeling level, defining a submodel by removing genes is equivalent to the effect of a gene knockout obtained from a molecular construct. It is likely that the metabolism that dominates a particular physiological state owes more to kinetic regulation than can be accounted for with the metabolism structure only. Moreover, it is very hard to ascertain that a gene is not expressed at all and even in this case, the absence of mRNA does not exclude that the protein is still present at a low concentration, as a remnant of the previous growth phase. The construction of a series of submodels followed by their rescue is essentially a way to circumvent the hard constraint caused by model viability and exploit submodel properties that would be out of range.

The fitness function is the key component of our analysis. It helps to deduce the gene sets in the framework of a metabolic network rather than using only a co-expression of genes that do not necessarily indicate the co-activities. The proposed measure of fitness showed its capacity to capture the health status of a submodel and thus some control on our construction, as well as produce biologically meaningful results. However, we are not claiming that the function investigated here is optimal. Several lines of improvement will be envisaged in future work: the formulation of the growth equation should be improved by taking into account more metabolites, the proposed weighing scheme is certainly suboptimal and other dynamics properties of the model could be considered apart from the score derived from the LP-based minimization. In addition, the proposed enrichment score formulation is relatively *ad-hoc* and would need to be improved to accommodate an upgraded fitness function.

Materials and Methods

1. Genome-scale metabolic network

FBA, FVA, dead-end detection, reaction and gene knock-out analyses were performed with the *sybil* R package (v2.0.0) (Gelius-Dietrich *et al*, 2013). The *clp* solver (v1.16.10) with *inibarrier* algorithm via *clpAPI* R package (v1.2.7) (Gelius-Dietrich, 2016) was used for the simulation. The *sybil* tolerance was set to 1e-8.

The iMK735 model was adapted to hypoxic and normoxic environments. For normoxia, oxygen consumption was unrestricted as in the original model and took a value of 244 $mmol \cdot gDW^{-1} \cdot h^{-1}$ as given by the Minimum Total Flux algorithm where the sum of absolute values of fluxes was minimized. We then arbitrarily limited the available oxygen to 50 $mmol \cdot gDW^{-1} \cdot h^{-1}$ to simulate hypoxic conditions. Preliminary investigation showed that the model behavior did not depend too much on the exact value of this setting.

2. RNA-seq data

RNA-seq reads were aligned against the *Y. lipolytica* genome GCA_000002525.2_ASM252v1 using *tophat2* (v2.0.13) (Kim *et al*, 2013a). The number of reads mapped onto each gene locus was obtained with *HTSeq-count* (v0.6.1) (Anders *et al*, 2015). The RNA-seq read counts are normalized using the TMM method from the R Bioconductor package *edgeR* (v3.12.1) (Robinson & Oshlack, 2010), and subsequently transformed to $\log_2$-counts with *voom* transformation (Law *et al*, 2014). The average $\log_2$-counts in each condition was computed and referred to as **expr** in our analysis. RPKM expression denoted normalized expression by gene length and library size and was computed as number of reads per kilobase per million mapped reads.

Then with the R Bioconductor package *limma* (v3.30.13) (Smyth, 2005), we built a linear model with one factor per condition on the transformed data, and performed differential analysis via moderated *t*-statistics on 21 pair-wise contrasts between the seven conditions studied in Maguire *et al*.

3. GO term enrichment analysis

The analyses were done using R (v3.2.2). For comparison of enrichment methods, metaboGSE used *GO142* as input. topGO (v2.24.0) was performed independently with *classic* and *weight01* algorithms with *fisher* statistic (Alexa & Rahnenführer, 2016) using the GO annotation and gene universe as above and *nodeSize* = 5. Twenty-one sets of differentially expressed genes (cut-offs: fold-change = 2 and FDR = 0.05) in 21 pair-wise contrasts as mentioned above were studied. Gene set enrichment analysis was performed with the GSEA software (v2.2.3) (Subramanian *et al*, 2005) using the GO annotation and gene universe as above. The gene set of iMK735 with its RNA-seq transformed count data was provided as input, and all the 21 pair-wise contrasts were studied.

The obtained enriched GO terms among *GO142* containing between 3 and 50 genes from the model were considered for each method. *P*-values obtained with metaboGSE, topGO and GSEA were adjusted using Benjamini-Hochberg correction (FDR) (Benjamini & Hochberg, 1995), and adjusted *p*-values < 0.05 were considered significant. This adjustment was unnecessary for Maguire *et al*. results as none of their GO terms among *GO142* were enriched with a reported raw *p*-value < 0.05.

4. Statistical testing

Kruskal-Wallis (Kruskal & Wallis, 1952) test was applied as a non-parametric test for different distributions between multiple samples.

Acknowledgements

This work was supported by the Swiss National Science Foundation grant CRSII3_141848 (Sinergia) to Dominique Sanglard, Salomé Leibundgut and Marco Pagni. This work was also supported in part by the SystemsX.ch initiative under the projects MetaNetX and HostPathX. The computations were performed at the Vital-IT Center for high-performance computing (https://www.vital-it.ch) of the SIB Swiss Institute of Bioinformatics. SIB receives financial supports from the Swiss Federal Government through the State Secretariat for Education and Research (SER). We thank Ioannis Xenarios, Alan Bridge, Mark Ibberson, Frédéric Burdet, Mathias Ganter, and Jörg Stelling for helpful discussions and supports at different stages of the project, Olivier Martin for supports in MetaNetX development, and Claus Jonathan Fritzemeier for supports with *sybil*.

Author contributions

MP and VDTT designed and developed the algorithms. VDTT and MP performed the computational analyses. MP, VDTT, DS, ATC, and SA-V interpreted biological results. SM, MP, and VDTT cleaned the *Y. lipolytica* model iMK735. MP and VDTT wrote the manuscript. All the authors read and approved the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

REFERENCES

Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I & Nielsen J (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput. Biol.* **8:** e1002518

Akesson M, Förster J & Nielsen J (2004) Integration of gene expression data into genome-scale metabolic models. *Metab. Eng.* **6:** 285–293

Alexa A & Rahnenführer J (2016) topGO: Enrichment Analysis for Gene Ontology. R package version 2.24.0. Available at: https://doi.org/doi:10.18129/B9.bioc.topGO

Anders S, Pyl PT & Huber W (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31:** 166–169

Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ & Herrgard MJ (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat. Protoc.* **2:** 727–738

Becker SA & Palsson BO (2008) Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.* **4:** e1000082

Benjamini Y & Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57:** 289–300

van Berlo RJP, de Ridder D, Daran J-M, Daran-Lapujade PAS, Teusink B & Reinders MJT (2011) Predicting metabolic fluxes using gene expression differences as constraints. *IEEEACM Trans. Comput. Biol. Bioinforma. IEEE ACM* **8:** 206–216

Brandes A, Lun DS, Ip K, Zucker J, Colijn C, Weiner B & Galagan JE (2012) Inferring carbon sources from gene expression profiles using metabolic flux models. *PloS One* **7:** e36947

Chandrasekaran S & Price ND (2010) Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci.* **107:** 17845–17850

Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, Farhat MR, Cheng T-Y, Moody DB, Murray M & Galagan JE (2009) Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. *PLoS Comput. Biol.* **5:** e1000489

Collins SB, Reznik E & Segrè D (2012) Temporal expression-based analysis of metabolism. *PLoS Comput. Biol.* **8:** e1002781

Fang X, Wallqvist A & Reifman J (2012) Modeling phenotypic metabolic adaptations of Mycobacterium tuberculosis H37Rv under hypoxia. *PLoS Comput. Biol.* **8:** e1002688

Feist AM & Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat. Biotechnol.* **26:** 659–667

Gelius-Dietrich G (2016) clpAPI: R Interface to C API of COIN-OR Clp. R package version 1.2.7. Available at: http://CRAN.R-project.org/package=clpAPI

Gelius-Dietrich G, Desouki AA, Fritzemeier CJ & Lercher MJ (2013) sybil – Efficient constraint-based modelling in R. *BMC Syst. Biol.* **7:** 125

Henry CS, Broadbelt LJ & Hatzimanikatis V (2007) Thermodynamics-Based Metabolic Flux Analysis. *Biophys. J.* **92:** 1792–1805

Imam S, Schäuble S, Brooks AN, Baliga NS & Price ND (2015) Data-driven integration of genome-scale regulatory and metabolic network models. *Front. Microbiol.* **6:** 409

Jensen PA & Papin JA (2011) Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics* **27:** 541–547

Kavšček M, Bhutada G, Madl T & Natter K (2015) Optimization of lipid production with a genome-scale model of Yarrowia lipolytica. *BMC Syst. Biol.* **9:** 72

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R & Salzberg SL (2013a) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14:** R36

Kim HU, Kim WJ & Lee SY (2013b) Flux-coupled genes and their use in metabolic flux analysis. *Biotechnol. J.* **8:** 1035–1042

Kim J & Reed JL (2012) RELATCH: relative optimality in metabolic networks explains robust metabolic and regulatory responses to perturbations. *Genome Biol.* **13:** R78

Kim MK, Lane A, Kelley JJ & Lun DS (2016) E-Flux2 and SPOT: Validated Methods for Inferring Intracellular Metabolic Flux Distributions from Transcriptomic Data. *Plos One* **11:** e0157101

Kim MK & Lun DS (2014) Methods for integration of transcriptomic data in genome-scale metabolic models. *Comput. Struct. Biotechnol. J.* **11:** 59–65

Kruskal WH & Wallis WA (1952) Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **47:** 583–621

Law CW, Chen Y, Shi W & Smyth GK (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15:** R29

Ledesma-Amaro R & Nicaud J-M (2016) Yarrowia lipolytica as a biotechnological chassis to produce usual and unusual fatty acids. *Prog. Lipid Res.* **61:** 40–50

Lee D, Smallbone K, Dunn WB, Murabito E, Winder CL, Kell DB, Mendes P & Swainston N (2012) Improving metabolic flux predictions using absolute gene expression data. *BMC Syst. Biol.* **6:** 73

Lex A, Gehlenborg N, Strobelt H, Vuillemot R & Pfister H (2014) UpSet: Visualization of Intersecting Sets. *IEEE Trans. Vis. Comput. Graph. IEEE InfoVis14*

Machado D & Herrgård M (2014) Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. *PLOS Comput Biol* **10:** e1003580

Maguire SL, Wang C, Holland LM, Brunel F, Neuvéglise C, Nicaud J-M, Zavrel M, White TC, Wolfe KH & Butler G (2014) Zinc finger transcription factors displaced SREBP proteins as the major Sterol regulators during Saccharomycotina evolution. *PLoS Genet.* **10:** e1004076

Mahadevan R & Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5:** 264–276

McCloskey D, Palsson BØ & Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol. Syst. Biol.* **9:** 661

Moretti S, Martin O, Van Du Tran T, Bridge A, Morgat A & Pagni M (2016) MetaNetX/MNXref - reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.* **44:** D523-526

Moxley JF, Jewett MC, Antoniewicz MR, Villas-Boas SG, Alper H, Wheeler RT, Tong L, Hinnebusch AG, Ideker T, Nielsen J & Stephanopoulos G (2009) Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator Gcn4p. *Proc. Natl. Acad. Sci. U. S. A.* **106:** 6477–6482

Navid A & Almaas E (2012) Genome-level transcription data of Yersinia pestis analyzed with a new metabolic constraint-based approach. *BMC Syst. Biol.* **6:** 150

Oberhardt MA, Palsson BØ & Papin JA (2009) Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* **5:** 320

O'Brien EJ, Monk JM & Palsson BO (2015) Using Genome-scale Models to Predict Biological Capabilities. *Cell* **161:** 971–987

Rébora K, Laloo B & Daignan-Fornier B (2005) Revisiting Purine-Histidine Cross-Pathway Regulation in Saccharomyces cerevisiae. *Genetics* **170:** 61–70

Robaina Estévez S & Nikoloski Z (2015) Context-Specific Metabolic Model Extraction Based on Regularized Least Squares Optimization. *PloS One* **10:** e0131875

Robinson MD & Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11:** R25

Rossell S, Huynen MA & Notebaart RA (2013) Inferring Metabolic States in Uncharacterized Environments Using Gene-Expression Measurements. *PLOS Comput Biol* **9:** e1002988

Schmidt BJ, Ebrahim A, Metz TO, Adkins JN, Palsson BØ & Hyduke DR (2013) GIM3E: condition-specific models of cellular metabolism developed from metabolomics and expression data. *Bioinforma. Oxf. Engl.* **29:** 2900–2908

Sergushichev AA, Loboda AA, Jha AK, Vincent EE, Driggers EM, Jones RG, Pearce EJ & Artyomov MN (2016) GAM: a web-service for integrated transcriptional and metabolic network analysis. *Nucleic Acids Res.* **44:** W194-200

Simeonidis E & Price ND (2015) Genome-scale modeling for metabolic engineering. *J. Ind. Microbiol. Biotechnol.* **42:** 327–338

Smyth GK (2005) limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* pp 397–420. Springer, New York, NY

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES & Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102:** 15545–15550

Töpfer N, Jozefczuk S & Nikoloski Z (2012) Integration of time-resolved transcriptomics data with flux-based methods reveals stress-induced metabolic adaptation in Escherichia coli. *BMC Syst. Biol.* **6:** 148

UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.* **43:** D204-212

Varma A & Palsson BO (1994) Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Nat. Biotechnol.* **12:** 994–998

Vivek-Ananth RP & Samal A (2016) Advances in the integration of transcriptional regulatory information into genome-scale metabolic models. *Biosystems* **147:** 1–10

Wang Y, Eddy JA & Price ND (2012) Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst. Biol.* **6:** 153

Zur H, Ruppin E & Shlomi T (2010) iMAT: an integrative metabolic analysis tool. *Bioinforma. Oxf. Engl.* **26:** 3140–3142

FIGURES

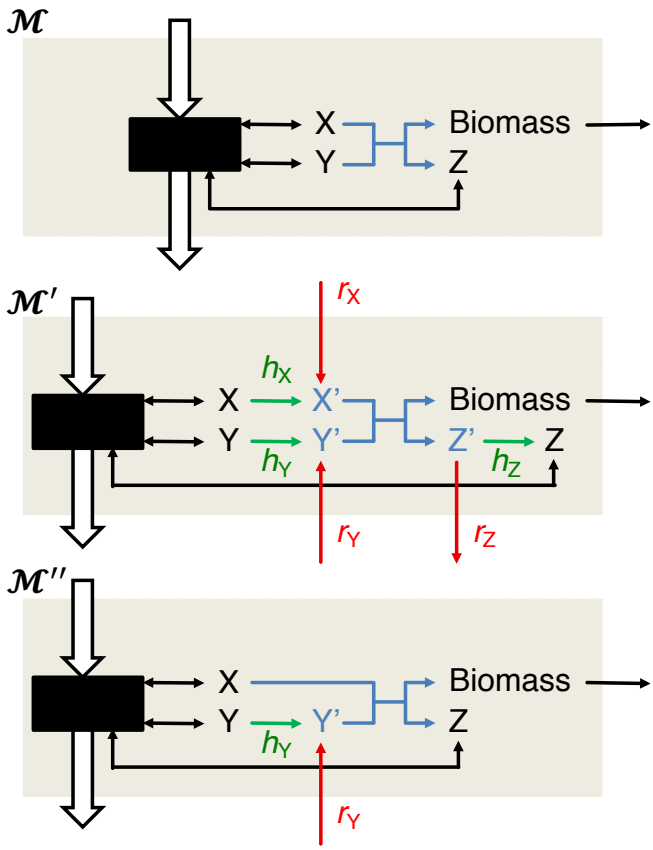


Figure 1. Schema of GSMN rescue process. $\mathcal{M}$, original GSMN with growth reaction X + Y ---> Z + Biomass. $\mathcal{M}'$, expanded GSMN with the full set of rescue (r_x) and help (h_x) reactions for every metabolite x in the biomass reaction. $\mathcal{M}''$, example of a minimal rescued GSMN in the particular case of only metabolite Y needs to be rescued.

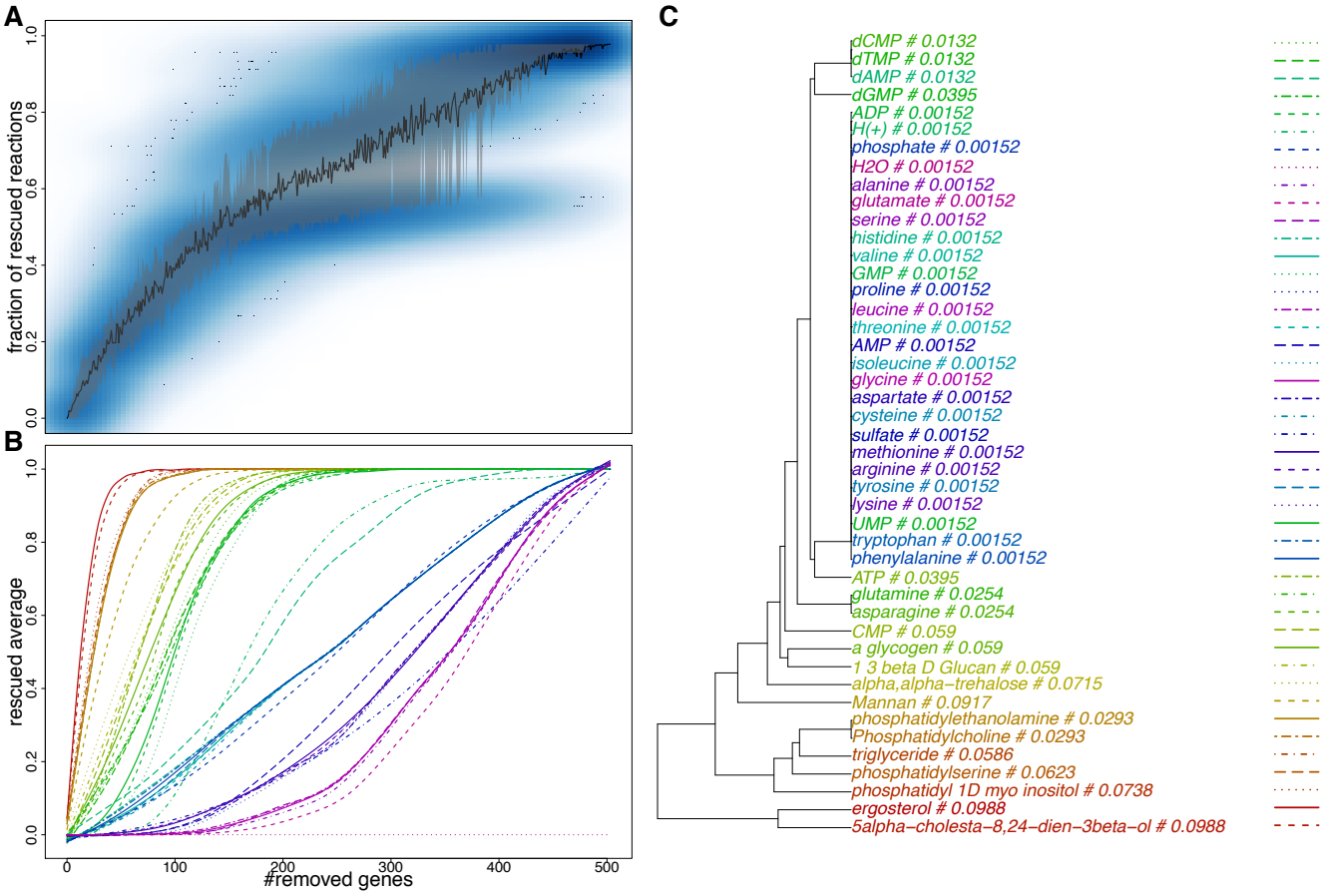


Figure 2. Rescue in metabolic submodels obtained by removing different numbers of random genes (*N* = 50 draws) from *Y. lipolytica* model iMK735.

(A) Fraction of metabolites in the growth reaction that need to be rescued. Blue darkness in the scatter plot represents the density of points. Black curves indicate average fitness.

(B) Fraction of draws where each individual metabolite is rescued.

(C) Weights of rescue reactions obtained via our weighting schema.

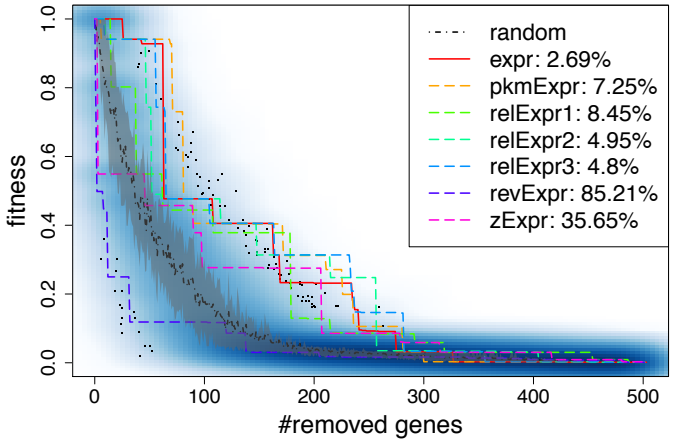


Figure 3. Fitness decrease while removing genes from the iMK735 model with different rankings in the UH condition. random: random draw, **expr**: voom-normalized expression, **pkmExpr**: voom-normalized expression in RPKM, **relExpr*i***: relative expression **($expr^i$/<expr>)$^{1/i}$**, **revExpr**: reverse expression, **zExpr**: z-score. Blue darkness in the scatter plot represents the density of random fitness. Gray region represents 20%-80% quantiles of random fitness.

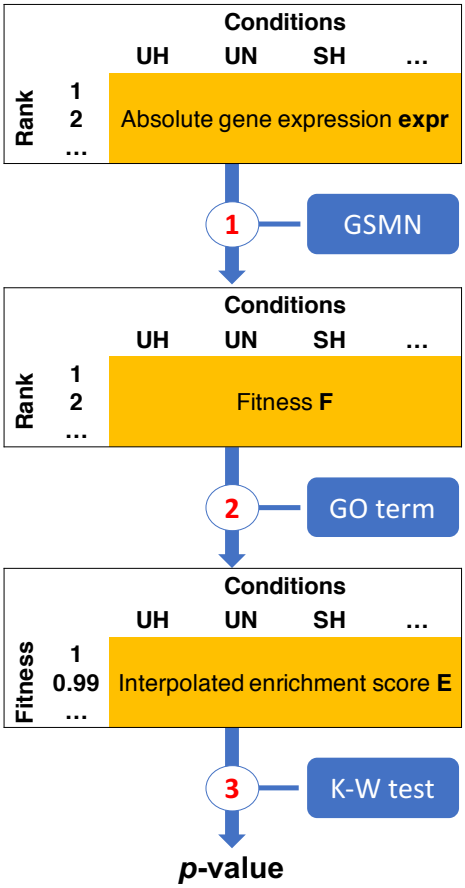


Figure 4. metaboGSE algorithm.

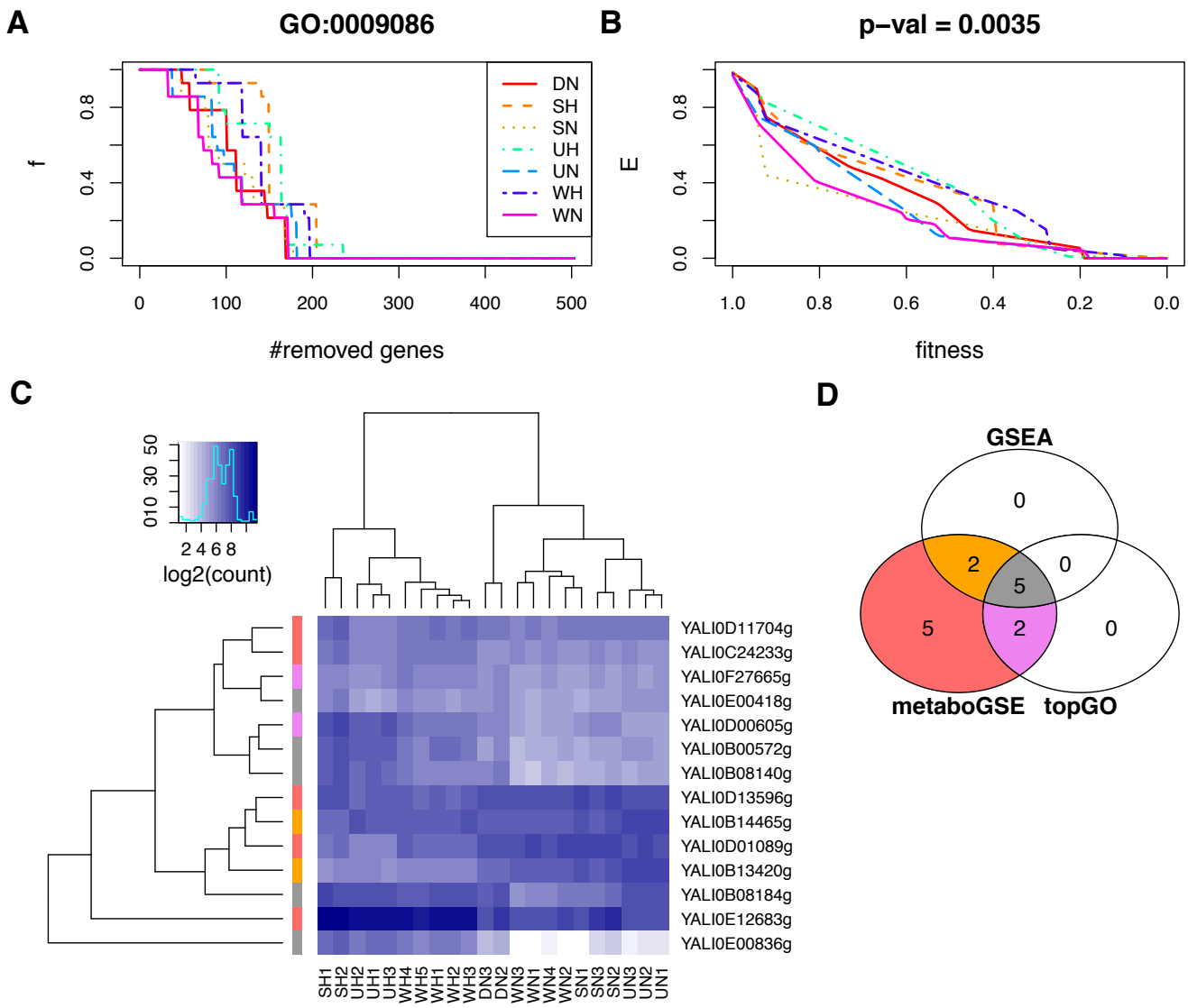


Figure 5. Enrichment of GO:0009086 methionine biosynthetic process in *Y. lipolytica* submodels in seven RNA-seq conditions.

(A) Fraction of remaining genes

(B) Enrichment score and Kruskal-Wallis *p*-value

(C) Expression of associated genes

(D) Venn diagram of associated genes studied in metaboGSE, GSEA, and topGO.

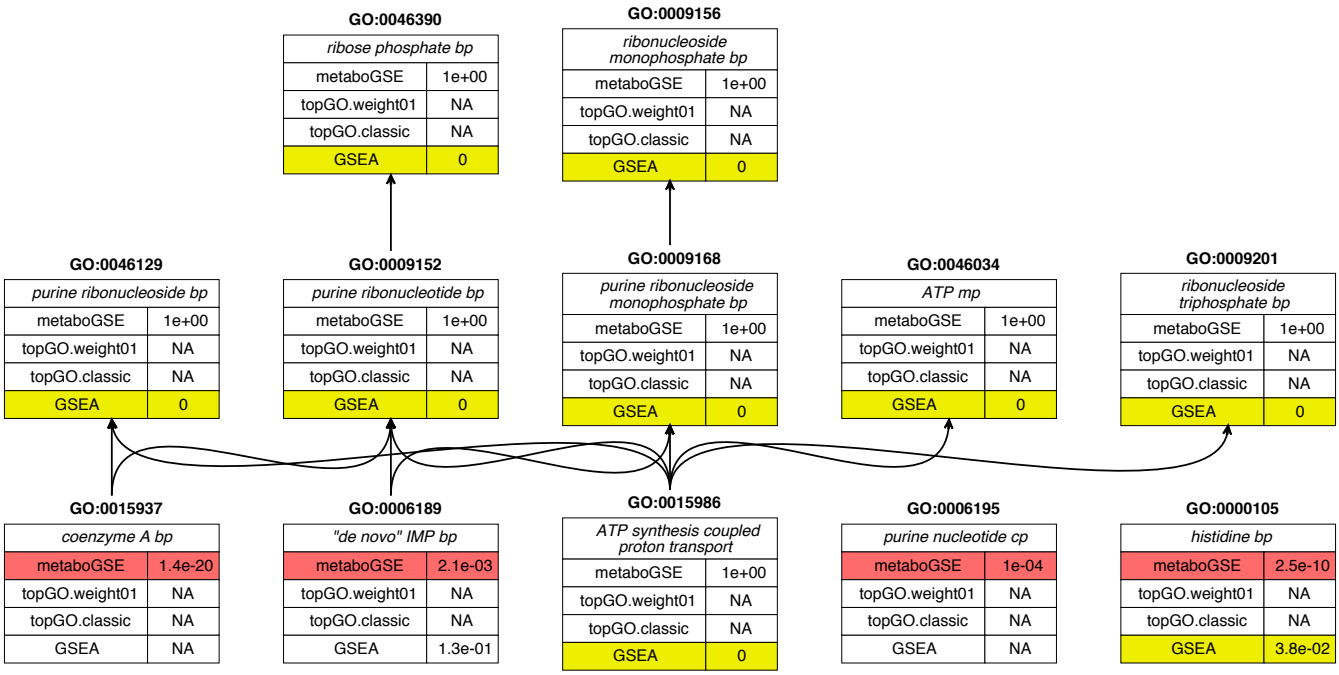


Figure 6. GO terms enriched via the three methods metaboGSE, GSEA, and topGO on the Purine-Histidine cross-pathway.

TABLES

Table 1. Designation of growth conditions obtained from Maguire *et al*.

#replicate	Code	Genotype	Oxygen
4	WN	wild type	normoxia 21%
3	SN	*sre1* Δ	
3	UN	*upc2* Δ	
2	DN	*sre1* Δ /*upc2* Δ	
5	WH	wild type	hypoxia 1%
2	SH	*sre1* Δ	
3	UH	*upc2* Δ	

Table 2. Performance of different rankings ρ measured by significance level Z_ρ. random: random draw, **expr**: voom-normalized expression, **pkmExpr**: voom-normalized expression in RPKM, **relExpr*i***: relative expression **(expr^i/<expr>)$^{1/i}$**, **revExpr**: reverse expression, **zExpr**: z-score.

	DN	**SH**	**SN**	**UH**	**UN**	**WH**	**WN**
expr	**3.27**	**7.91**	**2.54**	**2.69**	**2.49**	**4.30**	**2.91**
pkmExpr	4.83	12.49	3.40	7.25	5.60	7.00	3.71
relExpr1	29.09	15.12	12.70	8.45	12.74	6.00	10.99
relExpr2	7.36	12.11	2.98	4.95	3.39	5.77	5.27
relExpr3	5.68	12.77	3.46	4.80	5.39	6.36	4.29
revExpr	82.98	85.72	77.57	85.21	80.76	77.71	76.81
zExpr	40.05	45.29	12.74	35.65	16.95	28.19	45.64

EXPANDED VIEW

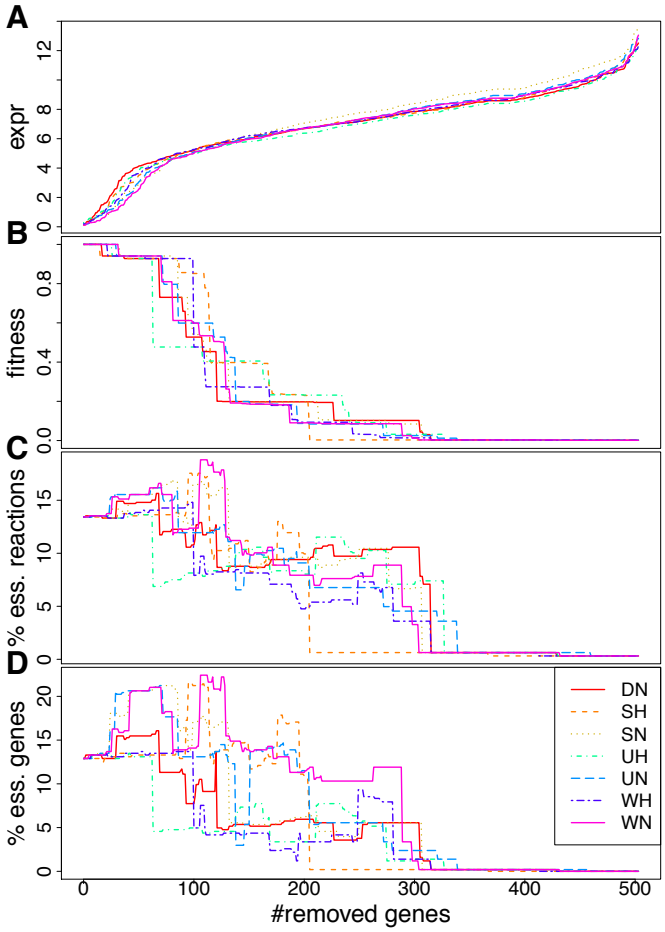


Figure EV1. Gene expression (A), fitness (B), percentage of essential reactions (C) and percentage of essential genes (D) of *Y. lipolytica* submodels from seven RNA-seq conditions, obtained by removing genes following the **expr** ranking. The percentages represent the numbers of essential reactions or genes in submodels over those in the comprehensive model. The region of stable fitness is up to a cut-off at 12.5% of the whole model gene set (63 genes).

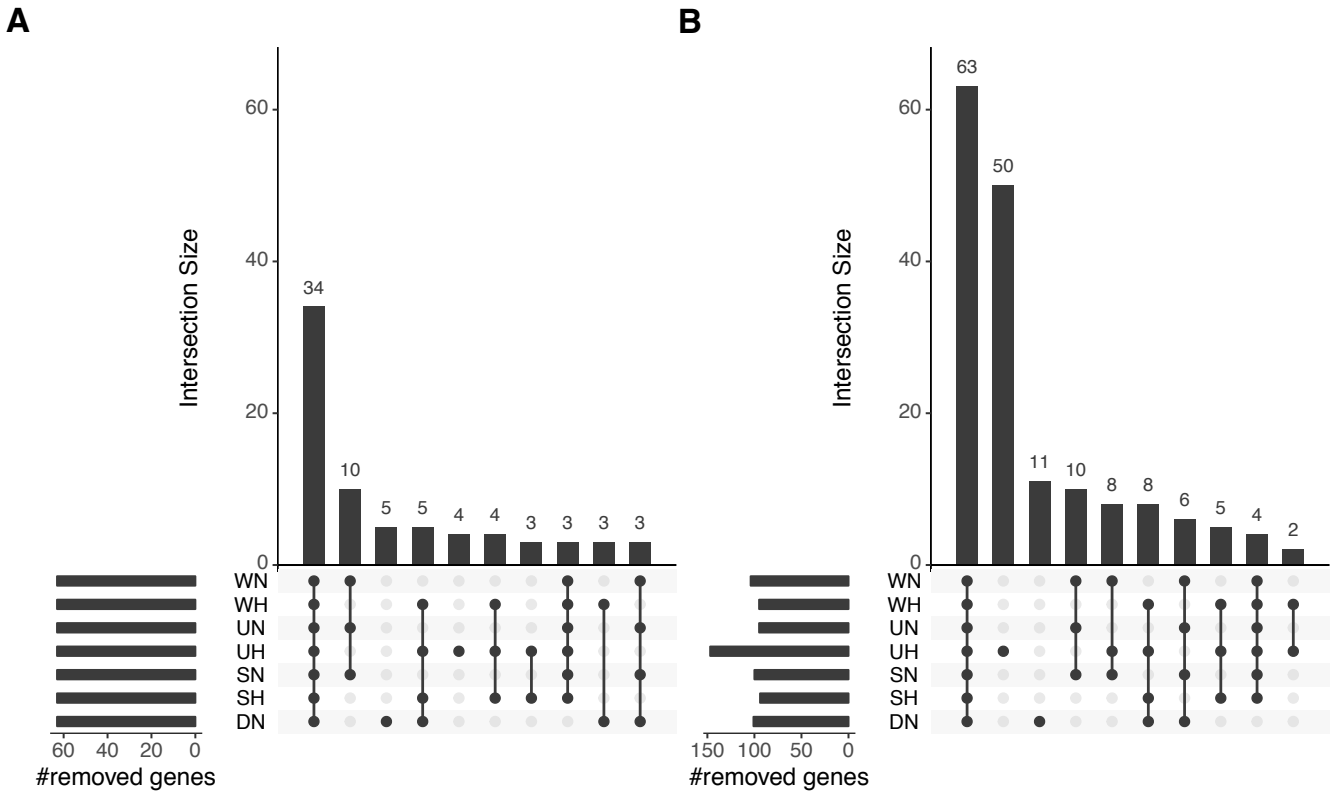


Figure EV2. Summary of genes removed in seven RNA-seq conditions.
A 12.5% (63) genes are removed following the **expr** ranking.
B Induced genes from removing 63 genes are also removed.

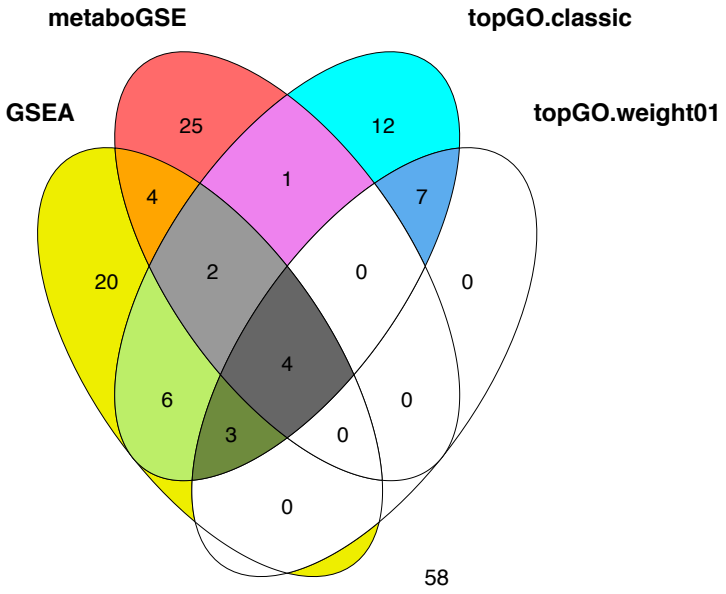


Figure EV3. GO terms enriched via the three methods metaboGSE, GSEA and topGO (with *classic* and *weight01* algorithms). FDR cut-off: 0.05.

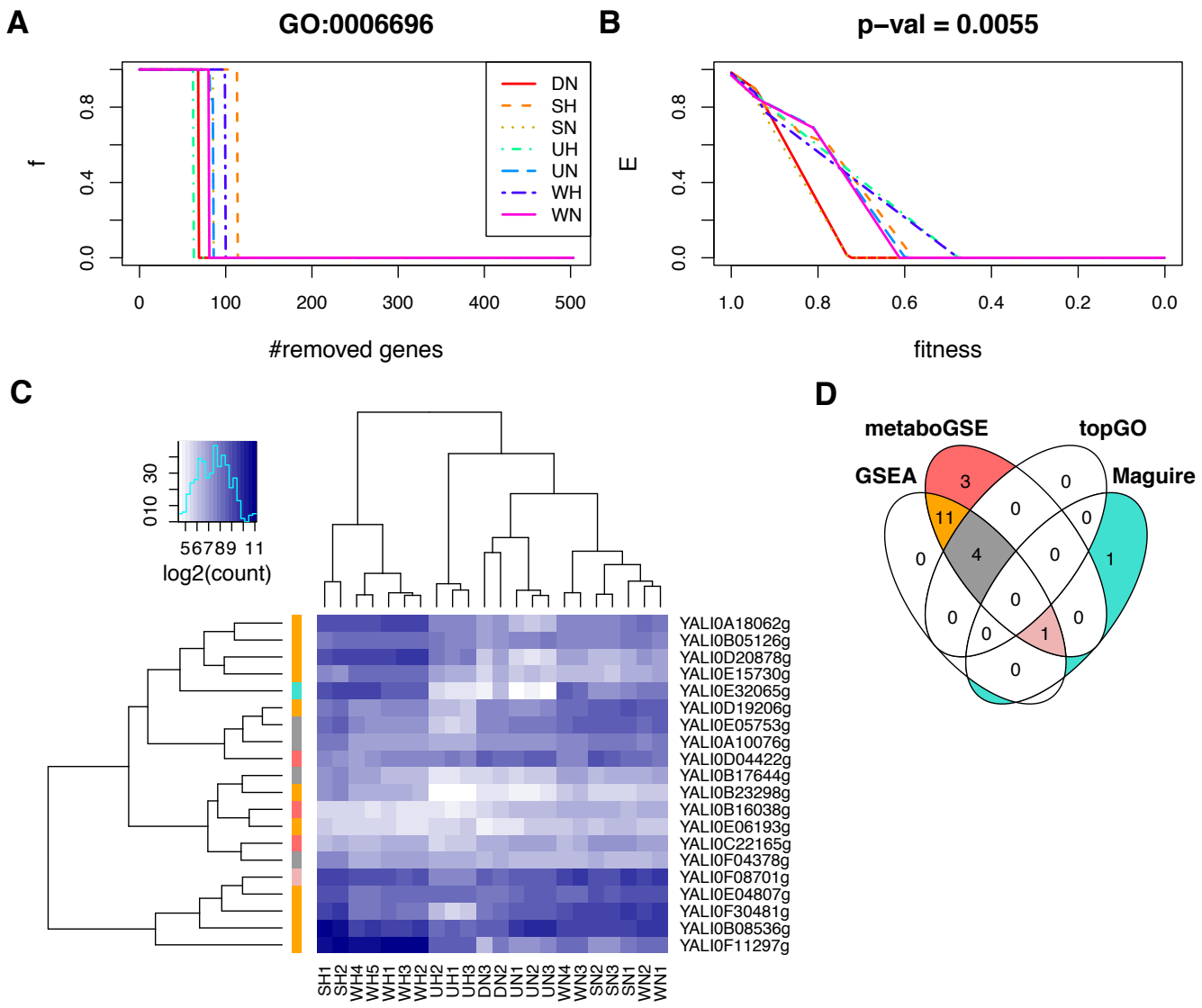
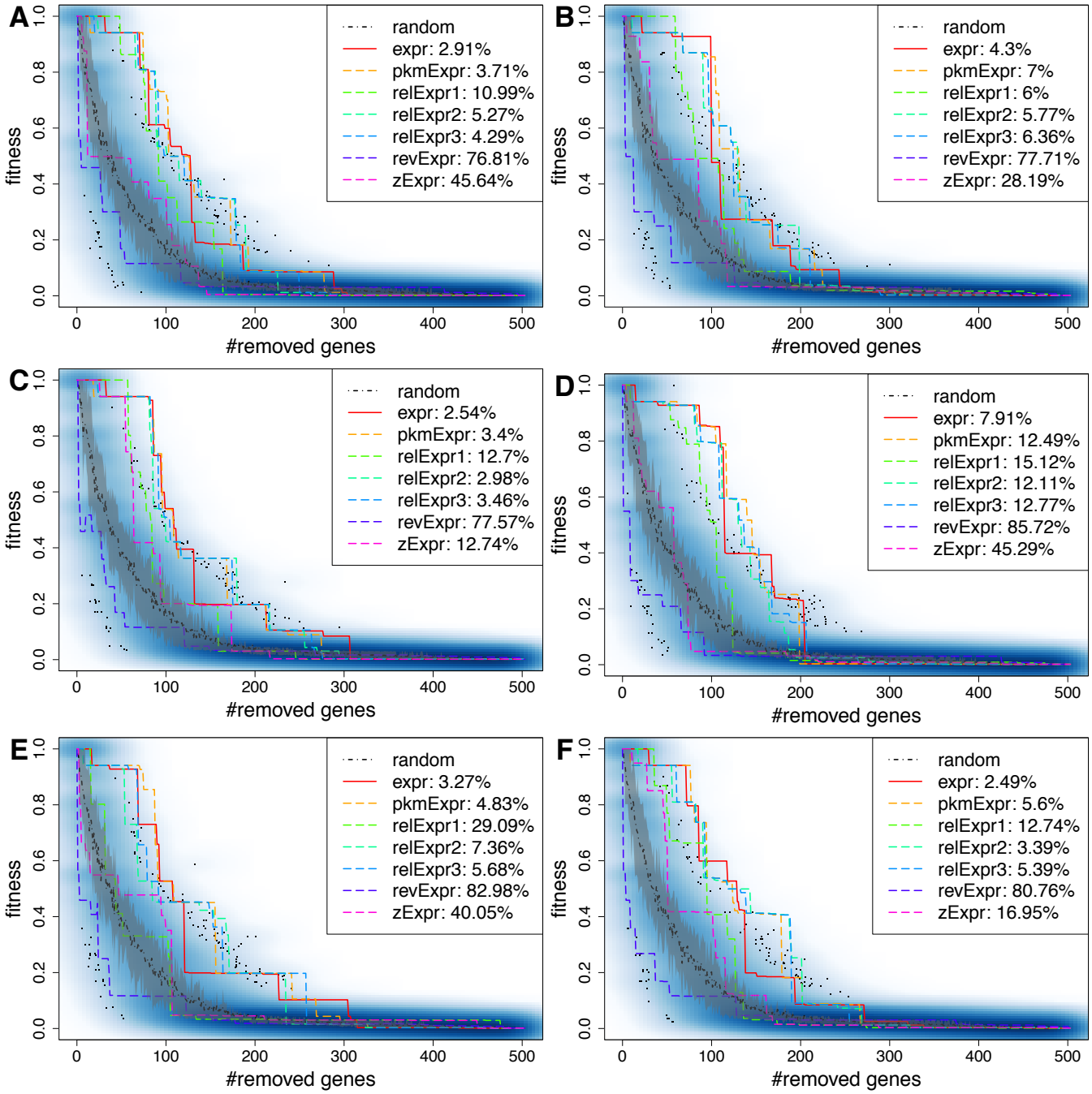


Figure EV4. Enrichment of GO:0006696 ergosterol biosynthetic process in *Y. lipolytica* submodels in seven RNA-seq conditions.

(A) Fraction of remaining genes
(B) Enrichment score
(C) Expression of ergosterol-associated genes
(D) Venn diagram of genes associated with GO:0006696 ergosterol biosynthetic process studied in metaboGSE, GSEA, topGO and genes associated with GO:0016126 sterol biosynthetic process reported in Maguire *et al*.

Table EV1. List of enriched biological process GO terms based on integration of RNA-seq data and metabolic networks compared to those obtained in Maguire *et al*. BH-adjusted p-values from Kruskal-Wallis test (metaboGSE), Fisher test (topGO), phenotype-based permutation test (GSEA) are cut off at 0.05. All biological process GO terms in *GO142* from Maguire *et al* with BH-corrected *p*-values are reported.

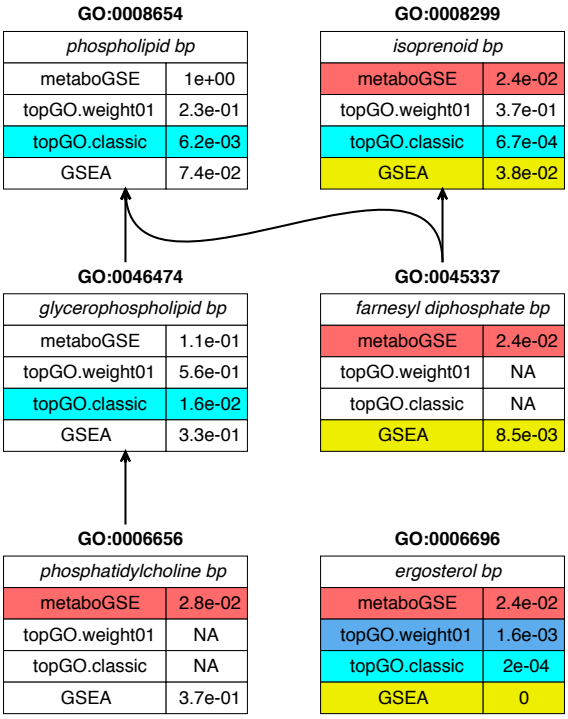
APPENDIX



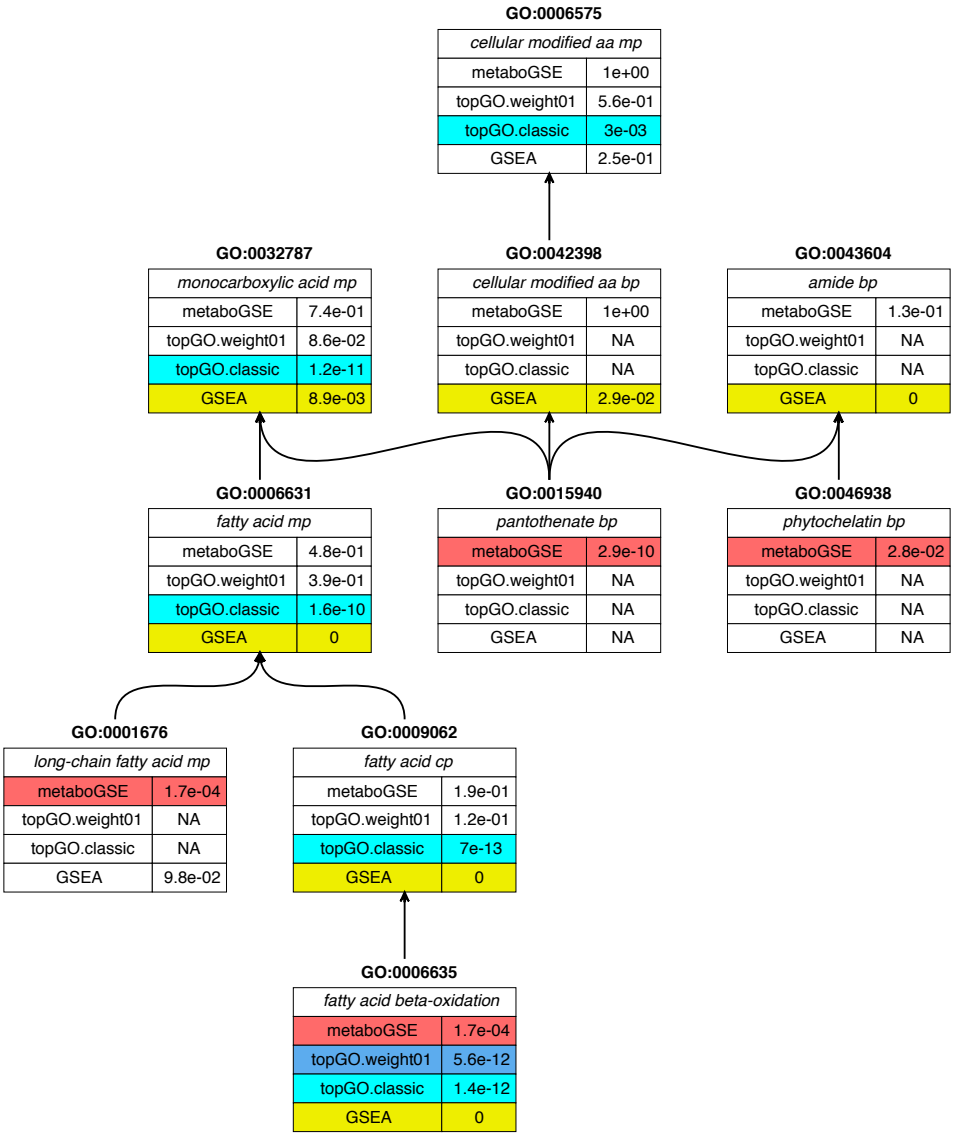
Appendix Figure S1. Fitness decrease while removing genes from the iMK735 model with different rankings. A. WN, B. WH, C. SN, D. SH, E. DN, F. UN. **random**: random draw, **expr**: voom-normalized expression, **pkmExpr**: voom-normalized expression in RPKM, **relExpr*i***: relative expression **($expr^i$/<expr>)$^{1/i}$**, **revExpr**: reverse expression, **zExpr**: z-score. Blue darkness in the

scatter plot represents the density of random fitness. Gray region represents 20%-80% quantiles of random fitness.

Appendix Figure S2. Fraction of remaining genes and enrichment score of 36 GO terms identified via metaboGSE in *Y. lipolytica* submodels in seven RNA-seq conditions.



Appendix Figure S3. GO terms enriched via the three methods metaboGSE, GSEA, and topGO on the ergosterol biosynthesis pathway.



Appendix Figure S4. GO terms enriched via the three methods metaboGSE, GSEA, and topGO on fatty acid metabolism.