Condition-specific series of metabolic sub-networks and its application for gene set enrichment analysis

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

* Correspondence: marco.pagni@sib.swiss

Running title: Gene set enrichment on metabolic network

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

SUMMARY

Genome-scale metabolic networks and transcriptomic data represent complementary sources of knowledge about an organism's metabolism, yet their integration to achieve biological insight remains challenging. We investigate here condition-specific series of metabolic sub-networks constructed by successively removing genes from a comprehensive network. The optimal order of gene removal is deduced from transcriptomic data. The sub-networks are evaluated via a fitness function, which estimates their degree of alteration. We then consider how a gene set, *i.e.* a group of genes contributing to a common biological function, is depleted in the series of sub-networks to detect the difference between experimental conditions. The method, named *metaboGSE*, is validated on public data for *Yarrowia lipolytica*. It is shown to produce more specific GO terms compared to GSEA or topGO, and better discrimination between conditions compared to GIMME. Our sub-network construction method could also use proteomic or metabolomic data, and should be relevant to other applications.

INTRODUCTION

The advent of high throughput sequencing techniques, especially RNA sequencing, has greatly facilitated the experimental investigation of an organism's transcriptome under different physiological conditions. RNA-seq data consists of reads mapped onto an annotated genome, which permits quantitation of the transcript abundance of all predicted genes. These values can be used as a proxy to quantify gene expression or may provide hints about protein abundance 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 (Conesa et al., 2016). 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 process, function or sub-cellular location. GO terms are hierarchically arranged in a directed acyclic graph (DAG) whose structure can be explored by computational methods such as topGO (Alexa and Rahnenführer, 2016). Another popular method, GSEA (Subramanian et al., 2005), considers two-condition expression profiles and attempts to identify enriched sets of genes by investigating their distribution in a particular ordering of all genes, which is based on the correlation between expression and experimental condition.

Annotated genomes can also be used for 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 and Palsson, 2008; McCloskey et al., 2013; Oberhardt et al., 2009). These are complex biological networks that include thousands of interconnected nodes of four main types: metabolites; biochemical and transport reactions; enzymes and transporters; and genes. A GSMN can be turned into a 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 and 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 the essentiality of genes, for example, by simulating the effect on viability of a gene knockout, which can be validated against experimental data (Imam et al., 2015; O'Brien et al., 2015; Simeonidis and Price, 2015).

Methods for integrating RNA-seq data with GSMN can be classified in two broad categories: those aimed at predicting FBA flux distribution in the network and those aimed at extracting context-specific sub-networks (Kim and Lun, 2014; Machado and Herrgård, 2014; Opdam et al., 2017; Vijayakumar et al., 2017; Vivek-Ananth and Samal, 2016). The first category is reviewed and benchmarked by Machado and Herrgård (Machado and Herrgård, 2014), who concluded with two ambiguous observations: absolute and relative gene expression seem to perform similarly; and using gene expression as additional FBA constraints is not clearly superior to what is obtained with *parsimonious FBA*, an algorithm that does not depend on transcriptomic data at all. The second category, reviewed by Opdam et al. (Opdam et al., 2017), consists in removing genes whose products are supposed to be absent in a given condition. These sub-network extraction methods differ in the way they address the following questions: how many genes to remove; how to choose the genes to remove given the RNA-seq data; how to deal with unviable sub-models caused by the removal of essential genes. They yield a diversity of sub-models and it is still unclear whether one of these methods is really superior to the others. Moreover, the choice of cut-off for gene expression in their construction strongly affects the model content and predictive capacity (Opdam et al., 2017). To date, there has been little biological insight from contrasting such context-specific sub-networks, with the exception of the predictions and validations of the fluxes on a few metabolites of interest (Machado and Herrgård, 2014). The widely used method GIMME (Becker and Palsson, 2008) belongs to both categories. It creates context-specific sub-models that attempt to reconcile the preservation of model viability with various physiological data (gene expression, metabolic objective, etc.).

The method presented here considers a whole series of sub-networks built by successively removing genes from an initial comprehensive network, rather than a single, possibly optimal, sub-network. The optimal ranking of the genes to remove is determined as the most significant one when compared to random rankings, using gene expression data. The viability of any sub-network is ensured by the introduction of a minimal number of artificial reactions. This rescue procedure also allows the formulation of a *fitness* function that assesses how close an unviable sub-network is from a viable one. We then study how a gene set, *i.e.* a collection of genes contributing to a common biological function, is depleted in the series of sub-networks. The depletion curve is integrated with respect to the fitness and tested for statistical significance given variations among conditions and replicates.

To validate the method, we used public experimental data and a metabolic model for *Yarrowia lipolytica*, a yeast which is widely exploited in industrial microbiology for lipid production (Ledesma-Amaro and Nicaud, 2016). The metabolic model iMK735 of *Y. lipolytica* (Kavšček et al., 2015) was used, as it can simulate growth and production of lipids when oxygen consumption is limiting. Gene expression data were 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). The latter study showed that control of the sterol pathway was taken over by Upc2 in *Y. lipolytica* and in the *Saccharomycotina* group.

RESULTS

Sub-model construction by removing genes refers to the process of (i) removing an initial set of genes (input) and their associated reactions, (ii) determining the blocked reactions (see STAR Methods), (iii) removing the blocked reactions with their associated genes, if any. A similar

procedure named *deleteModelGenes* is available from the COBRA Toolbox (Heirendt et al., 2017). Below, we refer to a particular sub-model with the number of initially removed genes, but all presented analysis results have been obtained after gene removal propagation through the blocked reactions.

Measure of metabolic model fitness

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

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 reaction as illustrated 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. *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 constraints are applied on the fluxes of the rescue and help reactions apart from the directions. In addition, the same network modifications are applied to the non-growth associated maintenance reaction that does not allow for a zero flux.

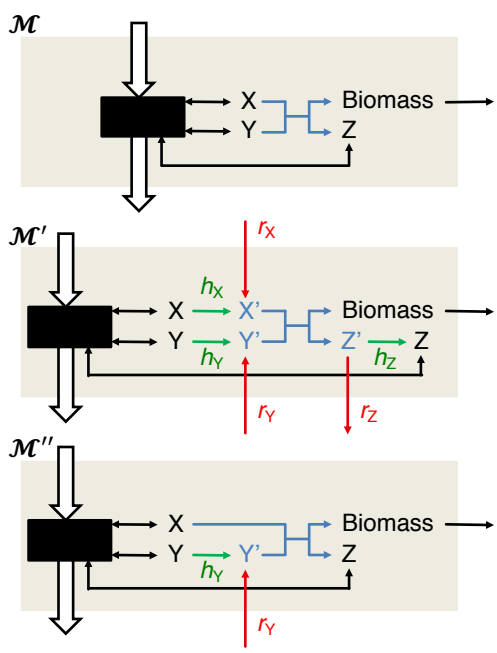

Figure 1. Schema of GSMN rescue process. $\mathcal{M}$, original GSMN with growth reaction X + Y ---> Z + Biomass. $\mathcal{M}'$, expanded GSMN with the full set of rescue (r_x) and help (h_x) reactions for every metabolite *x* in the biomass reaction. $\mathcal{M}''$, example of a minimal rescued GSMN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

Let $\mathcal{M}'$ be the expanded version of a GSMN $\mathcal{M}$ with the modified growth reaction and the full set of rescue and help reactions, $\mathbf{S}'$ is the corresponding stoichiometric matrix (see basic notions on GSMN in STAR Methods). The flux distribution $\boldsymbol{v}'$ is constrained by the bound vectors $\mathbf{lb}'$ and $\mathbf{ub}'$ that account for the directionality of newly added reactions and constraint the growth reaction 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 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 reaction. 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. A viable model $\mathcal{M}''$ with a minimal number of rescue reactions could be formulated as shown in Fig 1.

Fig 2A presents the fraction of metabolites in the growth reaction that need to be rescued after randomly removing genes. This simulation shows the crucial property that the more genes removed, the more damaged the model. Fig 2B presents the fraction of random draws where each individual metabolite needs to be rescued, clearly showing different behaviors among metabolites in the growth reaction. This suggests that the different metabolites should not be treated equivalently.

Weighting scheme for model fitness

A weighting scheme is introduced to account for the variable importance of metabolites in the growth reaction and possible dependencies among them (Fig 2B). Model fitness is defined as

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

where $\mathbb{I}(e) = 1$ if e is true, and 0 otherwise. The weights w_{r_x} are computed in the following procedure. 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 reaction are affected. Hence, every rescue reaction r_x can be associated with a binary vector which describes whether the reaction is needed to rescue 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 (Gerstein et al., 1994) 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 hardly 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 the same weight.

With the simulation of random draws above, the distribution of obtained sub-model fitness scores was less dispersed than that of the fraction of rescued metabolites (see Fig 2A and 3).

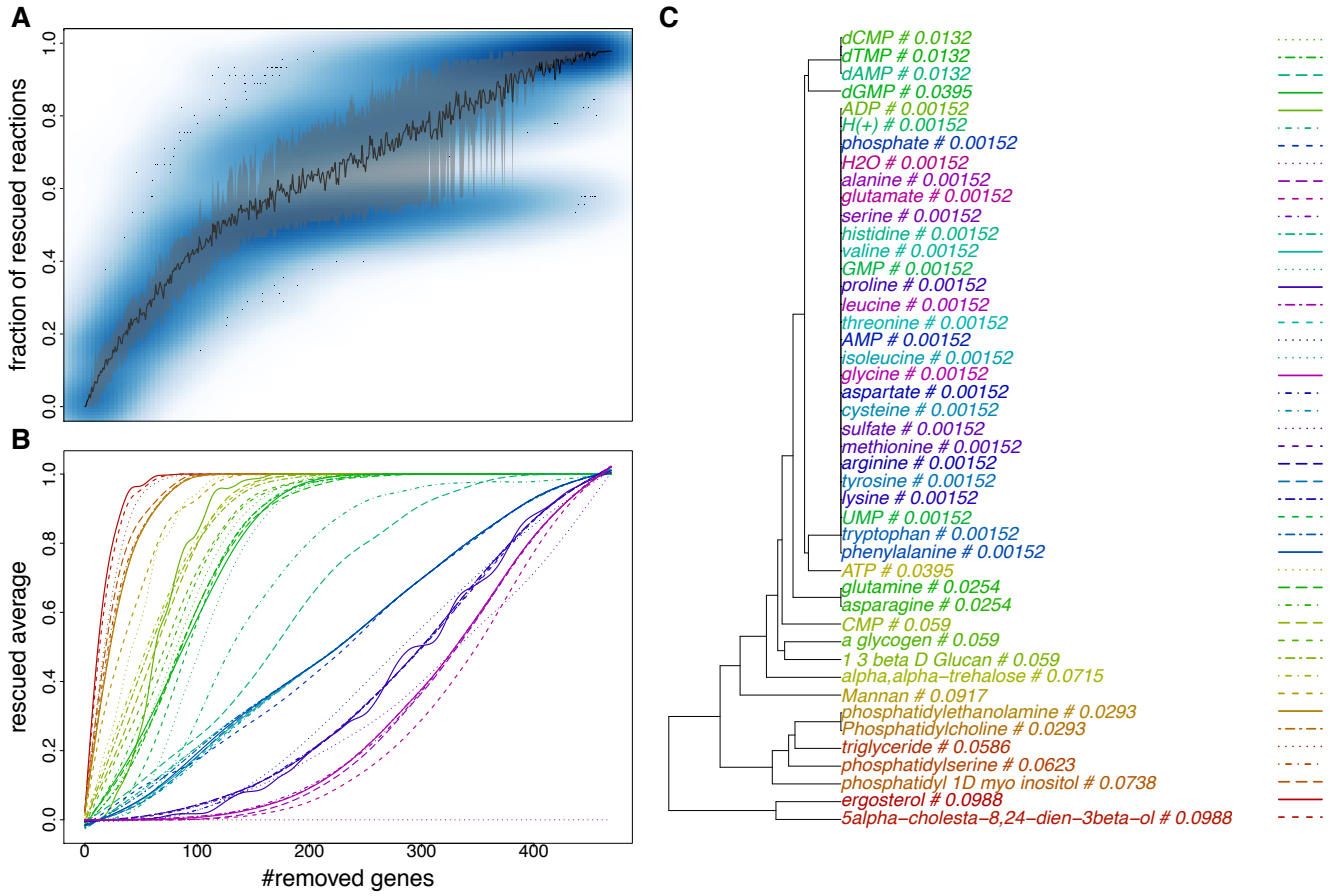

Figure 2. Rescue in metabolic sub-models obtained by removing different numbers of random genes (*N* = 50 draws) from *Y. lipolytica* model iMK735.
(A) Fraction of metabolites in the growth reaction that need to be rescued. Blue coloring in the scatter plot represents the density of points. Black curves indicate average fitness.
(B) Fraction of draws where each individual metabolite is rescued.
(C) Weights of rescue reactions obtained via our weighting schema.

Optimal ranking of genes for removal

The proposed fitness function can be used to evaluate a condition-specific metabolic sub-model constructed by removing a certain number of genes from the comprehensive network. The question then arises as to how to select genes for elimination in a particular condition, with the aim of removing as many genes as possible while minimizing network disruption, *i.e.* preserving its fitness. In this section, we investigate different metrics to rank genes for removal according to their expression.

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

expr	raw expression in $\log_2$-counts
pkmExpr	expression in RPKM (see STAR 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**)	reverse expression (control),

where <**expr**> and sd(**expr**) denote the average and standard deviation across conditions, respectively.

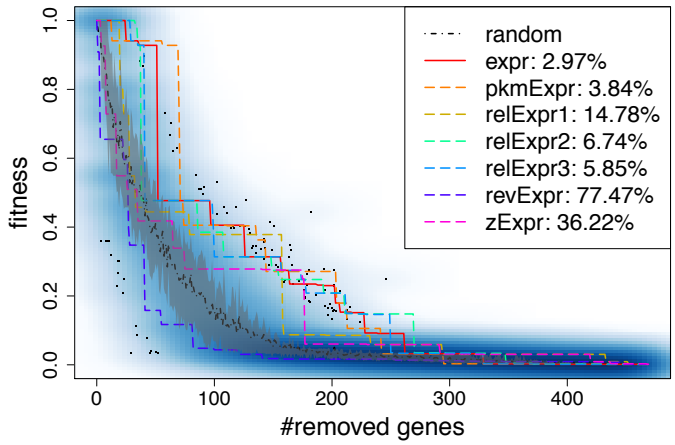

Figure 3. Fitness decrease while removing genes from the iMK735 model with different rankings in the UH condition (UH2 sample). random: random draw, **expr**: voom-normalized expression, **pkmExpr**: voom-normalized expression in RPKM, **relExpr*i***: relative expression **(expr^i/<expr>)$^{1/i}$**, **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.

Let ρ be a particular ranking of the genes. Sub-models are constructed by removing genes in the order given by ρ. The resulting fitness scores decrease when more genes are removed, and the decreasing trend depends on ρ. Fig 3 presents such reductions obtained by successively removing genes from the comprehensive model iMK735 for the UH condition (*upc2*Δ in hypoxic condition, UH2 sample) and for the seven rankings described above.

We assess a ranking ρ by *significance level* Z_ρ, which indicates the percentage of random draws yielding sub-model fitness higher than that of sub-models created by ρ-based removal of the same number of genes and which is weighted by the average random fitness.

Z_ρ is computed as follows:

$$Z_\rho = 100 \times \sum_i \left(\frac{\langle F\left(\mathrm{M}_i^{random}\right)\rangle}{\sum_j \langle F\left(\mathrm{M}_j^{random}\right)\rangle} \sum_{N \text{ draws}} \frac{1}{N} \mathbb{1}\left(F\left(\mathrm{M}_i^{random}\right) \geq F\left(\mathrm{M}_i^{\rho}\right)\right)\right),$$

where M_i^ρ denotes the rescued sub-model of the comprehensive network M by removing the first i genes following the ranking ρ, and M_i^{random} the rescued sub-model by randomly removing i genes, $\langle F(\ldots)\rangle$ the average fitness on N draws. The lower the significance level, the better the ranking. As shown in Fig 3 and Table S1, for this dataset, the absolute expression ranking **expr** dominates the others in producing the lowest significance level. The reverse expression ranking **revExpr**, where we remove genes with the highest expression first, was introduced as a control. Its high $Z_{\mathbf{revExpr}}$

value indicates that the reverse ranking is clearly worse than a random gene removal. Consequently, the **expr** ranking was then used in the rest of this study.

Condition-specific sub-model construction and comparison of multiple conditions

Now we apply the selected gene removal strategy to construct sub-models for the seven conditions with 22 RNA-seq samples of normoxic and hypoxic growth with *sre1*Δ, *upc2*Δ, *sre1*Δ /*upc2*Δ mutants and the wild type strain obtained from Maguire *et al* (NCBI: PRJNA205557) (see Table 1 and standard preliminary RNA-seq data analysis in STAR Methods). The **expr** ranking is specific to each condition despite a similar distribution of gene expression as depicted in Fig S1A (only one replicate per condition is shown). As a result, the sub-models constructed based on **expr**, and thus their fitness, evolve differently across conditions (Fig S1B). Fig S1C-F show a high degree of variation across conditions in the number of genes and reactions as well as in the proportion of essential genes and reactions, indicating that the corresponding sub-models are structurally distinct. Besides, the indistinct evolution of essential genes and reactions with regard to the number of removed genes reveal the unreliability when studying a unique condition-specific sub-model, as done in the existing methods, especially GIMME (Becker and Palsson, 2008).

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	

To compare the condition-specific sub-model construction from GIMME with that from our approach, we build a sequence of sub-models with GIMME using gene expression cut-offs (in RPKM) from 0 to 1000 for each of the 22 samples. Intersections and unions of genes in each 22 sub-models are investigated to evaluate their discrepancies. Fig S2 shows 1 - Intersection/Union plotted as a function of *k* - Union, where *k* denotes the number of genes in the comprehensive GSMN. Clearly our method yields more distant sub-models across conditions than GIMME. The latter method tends to a saturation which suggests that it is exploring a smaller sub-model space.

MetaboGSE: contrasting gene set enrichment in condition-specific sub-models

We introduce here the metaboGSE method, which identifies gene sets that are differentially enriched across conditions using a series of sub-models derived from a comprehensive GSMN. The method consists in three main steps depicted in Fig 4 and is illustrated in Fig 5 for GO:0006635 – fatty acid beta-oxidation. The first step consists in constructing a series of sub-models and is explained above.

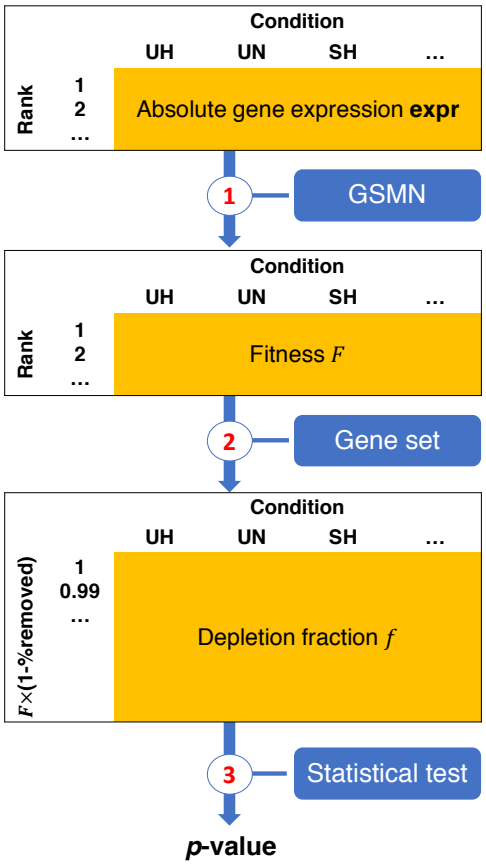

Figure 4. metaboGSE algorithm.

In the second step, with a given gene set g, we compute the *depletion fraction* $f(\mathrm{M}_i^{\mathbf{expr}}, g)$, *i.e.* fraction of g-associated genes remaining in each sub-model and for each library, where $\mathrm{M}_i^{\mathbf{expr}}$ denotes the rescued sub-model of the comprehensive network M by removing the first i genes. The evolution of f is studied as a function of $\left(1-\frac{i}{k}\right)\cdot F\left(\mathrm{M}_i^{\mathbf{expr}}\right)$, defining a depletion curve, where k denotes the total number of genes in M. The depletion curve for each condition is identified as the average curve from its replicates. As shown in Fig 5AB, the fraction f of genes associated with fatty acid beta-oxidation decreases rapidly in all conditions, while the depletion curve clearly separates hypoxic wild-type from the other conditions. The down-regulation of several, yet not all, GO:0006635 genes in the WH condition is also visible in Fig 5C. The genes that are not down-regulated are involved in the depletion due to their connectivity with the down-regulated ones.

The third step involves two statistical tests for the significance of the given gene set g. In the first randomization test, we investigate the significance of a gene set in a single experimental condition. For each condition, we generate depletion curves for random gene sets and compare their area under the curve (AUC) with that of g. The single p-value of g for each condition indicates the probability that AUC of g is higher than AUC of a random gene set. When studying multiple conditions, g is required to be significantly enriched in at least one condition. Fig 5D shows the significance of GO:0006635 in the hypoxic wild-type with its minimum p-value of 0.00059.

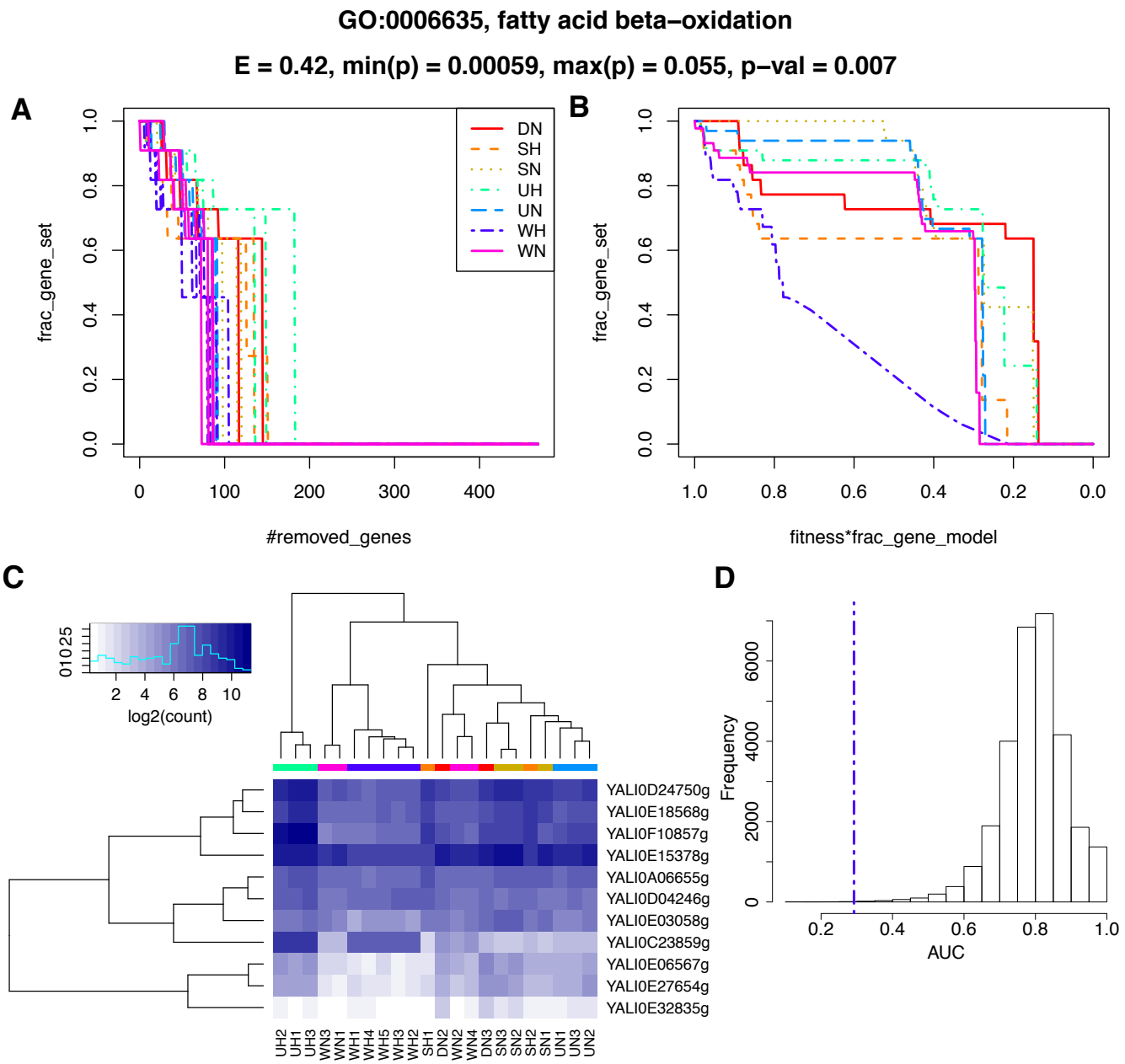

Figure 5. Enrichment of GO:0006635 fatty acid beta-oxidation in *Y. lipolytica* sub-models in seven RNA-seq conditions.

(A) Depletion fraction in function of number of removed genes

(B) Depletion curve: depletion fraction in function of fitness times fraction of remaining genes in the model

(C) Expression of associated genes

(D) AUC of the depletion curve in the hypoxic wild-type condition compares to AUC of random gene sets, producing the lowest single *p*-value min(p)

The second permutation test looks at the significance of discrepancy of g across conditions. The test statistic is defined as the maximum area between depletion curves of g from every pair of conditions. The resampling is performed by permuting replicates between conditions while keeping unchanged the number of replicates in each condition. The resulting p-value indicates whether the depletion evolution of g in at least one condition differ from those in the others. For GO:0006635, the discrepancy of WH versus the other conditions is justified with a p-value of 0.007 (Fig 5). These p-values are subsequently adjusted by Benjamini-Hochberg (BH) correction across all the studied gene sets (Benjamini and Hochberg, 1995). A similar post-hoc permutation test is also implemented for pairwise comparisons between conditions to check for the pairwise differential signal.

Biological interpretation

To validate the biological findings produced by our approach, we investigate in this section 135 gene sets defined as *GO135* (see STAR methods). Maguire *et al*. 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 *GO135*, but none has a reported BH-adjusted p-value < 0.05. Besides, no GO terms are significantly enriched from sub-models constructed by GIMME with its default parameters. Here we are comparing our results with those of topGO, using two algorithms *classic* and *weight01*, and GSEA (see STAR Methods).

Condition contrast is predominantly hypoxia-normoxia. Thirty-eight GO terms are found to be significantly enriched with metaboGSE, with FDR < 0.05 (see Table S2). Depletion curves (Fig S3) suggest that the degree of oxygen limitation is the most likely explanation for the enrichment of 21 GO terms, which is expected for this dataset – where hypoxia was the only environmental variable. Interestingly, for 16 of them, normoxic double-mutant condition is likely grouped with hypoxic conditions while contrasting with other normoxic conditions.

MetaboGSE reveals more specific GO terms. The GO terms found to be significantly enriched by metaboGSE, topGO, and GSEA are summarized in Fig S4 and Table S1. Of the 38 GO terms found by metaboGSE, 29 are related to those found by the other methods, yet include more specific terms. The largest connected DAG of enriched GO terms is illustrated in Fig 6. No terms from metaboGSE are parents of those found only by topGO and/or GSEA. All terms identified via metaboGSE and most of terms identified via topGO or GSEA are either children or unrelated of those reported in Maguire *et al*.

Difference in sterol biosynthesis was confirmed. Ergosterol biosynthetic process (GO:0006696) is enriched by all methods. Along with GSEA, we also discover GO:0045337 (farnesyl diphosphate biosynthetic process), which is part of the ergosterol synthesis pathway (see Fig S5). This reflects the experimental design investigated in Maguire *et al*., *i.e.* the regulation of sterol metabolism. In Fig S6AB, the sudden drop from 100% to 0% of ergosterol-associated genes indicates that either all these genes or none of them exist in each sub-model. This is due to the linearity of the pathway of ergosterol biosynthesis, which is not revealed by the expression profile of the genes (Fig S6C). metaboGSE detects the difference between conditions by the altered ranks of the lowest expression ergosterol-associated gene and its connectivity with other genes in the GSMN, unlike topGO and GSEA that are based on the differential expression of several genes associated with GO:006696. A tendency of hypoxia-normoxia contrast could be observed for these ergosterol biosynthesis-related GO terms (Fig S6B).

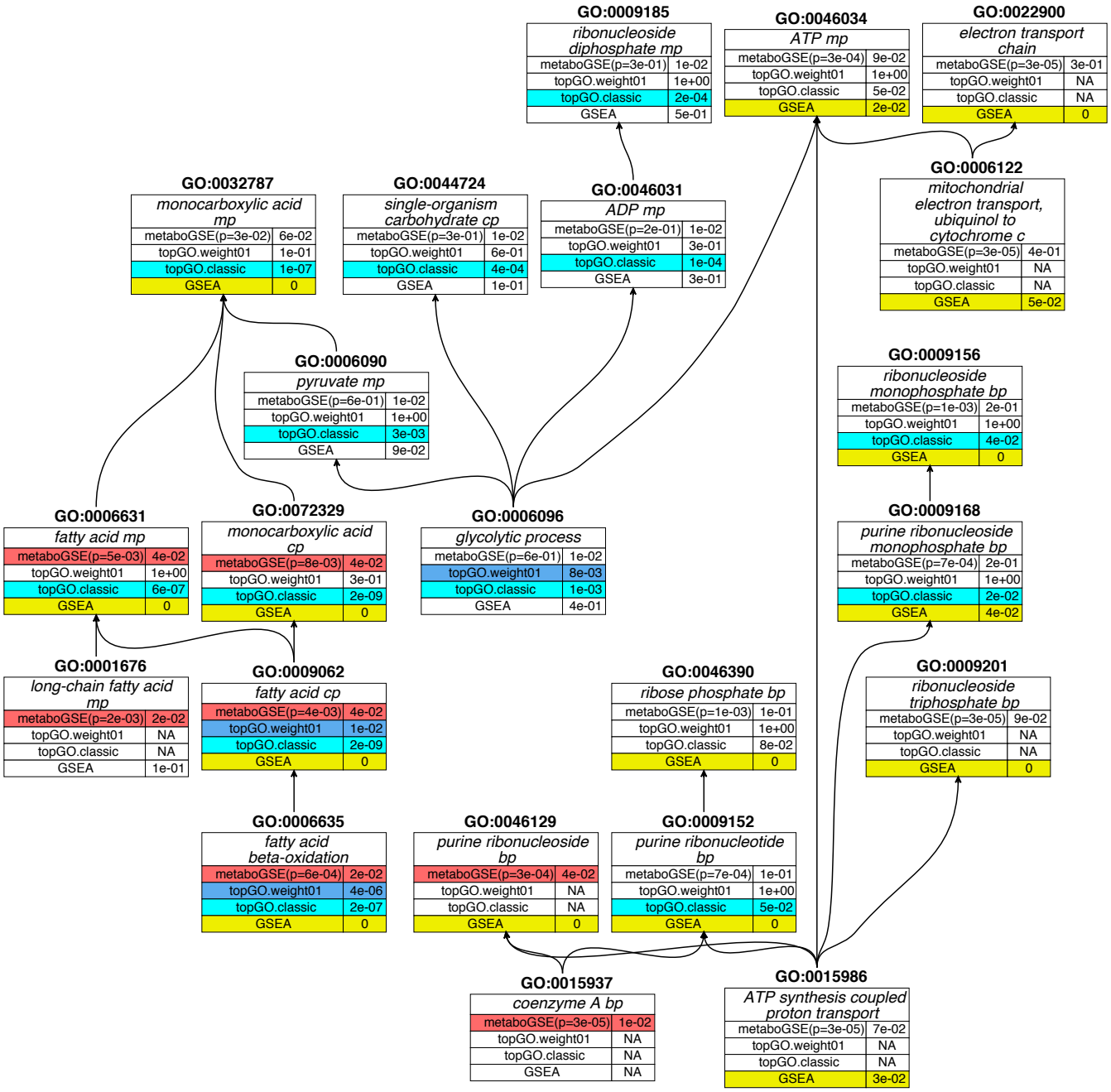

Figure 6. The largest connected DAG of GO terms enriched via the three methods metaboGSE, GSEA, and topGO. bp: biosynthetic process, cp: catabolic process, mp: metabolic process.

Contrasts other than hypoxia-normoxia have been identified with metaboGSE. All GO terms relating to fatty acid metabolism as well as the tryptophan biosynthetic process show an earlier drop in their depletion curve in the hypoxic wildtype comparing to the remaining conditions (see Fig 6 and FigS3). This might suggest an altered behavior of associated genes in the presence of both Sre1 and Upc2 in low-oxygen growth condition.

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 level. Our central working hypothesis is that the correlation between gene expression levels and fluxes on related reactions is very poor in general, but the low 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 detectably expressed genes. The introduction of a GSMN restricts the list of investigated genes to those present in the model (*i.e.* related to metabolism), and thus the list of gene sets that can be discovered. Clearly, the quality and scope of the tested 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 (*i.e.* that are located lower in the Gene Ontology hierarchy) 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 in the case reported here. The genes affecting the discrepancy between conditions, which are not necessarily differentially expressed, can be further investigated for each enriched gene set. 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 sub-model, but rather integrates on a series 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 sub-model by removing genes is equivalent to the effect of a gene knockout obtained from a molecular construct. It is likely that the metabolism dominating 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. Likewise, the presence of the protein does not ensure its activity. The construction of a series of sub-models followed by their rescue is essentially a way to circumvent the hard constraint caused by model viability and exploit sub-model properties that would be out of range. It appears to explore more profoundly differential signals between conditions compared to when studying a single condition-specific sub-model as in GIMME. Our method of construction of metabolic sub-networks could also be performed with other omics data, such as proteomics and metabolomics, and applied to other applications.

The fitness function is the key component of our method. It helps to deduce the gene sets in the framework of a metabolic network rather than using only the co-expression of genes that do not necessarily indicate co-activities. The proposed measure of fitness shows its capacity to capture the health status of a sub-model and thus suggests some control on our construction. Despite the biologically meaningful results obtained with the studied dataset, several lines of improvement can be envisaged in future work: the formulation of the growth reaction should be improved by taking into account more metabolites; the proposed weighting scheme is certainly suboptimal; and other dynamic properties of the model could be considered apart from the score derived from the LP-based minimization.

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, Mark Ibberson, Frédéric Burdet, Alan Bridge, Nicolas Guex, Brian Stevenson, Frédéric Schütz, 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

Conceptualization, MP and VDTT; Methodology: VDTT and MP; Software: VDTT and MP; Formal Analysis: VDTT and MP; Data Curation: VDTT, SM, and MP; Investigation: MP, VDTT, SM, DS, ATC, and SA-V; Writing – Original Draft: MP and VDTT; Writing – Review & Editing: MP, VDTT, SM, DS, ATC, and SA-V; Supervision: MP; Funding Acquisition: MP and DS.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Alexa, A., Rahnenführer, J., 2016. topGO: Enrichment Analysis for Gene Ontology. R package version 2.24.0.

Anders, S., Pyl, P.T., Huber, W., 2015. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166–169.

Becker, S.A., Palsson, B.O., 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.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X., et al., 2016. A survey of best practices for RNA-seq data analysis. Genome Biol. 17.

Feist, A.M., Palsson, B.Ø., 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.

Gelius-Dietrich, G., Desouki, A.A., Fritzemeier, C.J., Lercher, M.J., 2013. sybil – Efficient constraint-based modelling in R. BMC Syst. Biol. 7, 125.

Gerstein, M., Sonnhammer, E.L., Chothia, C., 1994. Volume changes in protein evolution. J. Mol. Biol. 236, 1067–1078.

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S.N., Richelle, A., Heinken, A., Haraldsdottir, H.S., Keating, S.M., Vlasov, V., Wachowiak, J., et al., 2017. Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0. ArXiv171004038 Q-Bio.

Henry, C.S., Broadbelt, L.J., Hatzimanikatis, V., 2007. Thermodynamics-Based Metabolic Flux Analysis. Biophys. J. 92, 1792–1805.

Imam, S., Schäuble, S., Brooks, A.N., Baliga, N.S., Price, N.D., 2015. Data-driven integration of genome-scale regulatory and metabolic network models. Front. Microbiol. 6, 409.

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.

Kersey, P.J., Allen, J.E., Armean, I., Boddu, S., Bolt, B.J., Carvalho-Silva, D., Christensen, M., Davis, P., Falin, L.J., Grabmueller, C., et al., 2016. Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res. 44, D574–D580.

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

Kim, M.K., Lun, D.S., 2014. Methods for integration of transcriptomic data in genome-scale metabolic models. Comput. Struct. Biotechnol. J. 11, 59–65.

Law, C.W., Chen, Y., Shi, W., Smyth, G.K., 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.

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, S.L., Wang, C., Holland, L.M., Brunel, F., Neuvéglise, C., Nicaud, J.-M., Zavrel, M., White, T.C., Wolfe, K.H., 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, C.H., 2003. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab. Eng. 5, 264–276.

McCloskey, D., Palsson, B.Ø., Feist, A.M., 2013. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol. Syst. Biol. 9, 661.

Moretti, S., Martin, O., Tran, V.D.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.

Oberhardt, M.A., Palsson, B.Ø., Papin, J.A., 2009. Applications of genome-scale metabolic reconstructions. Mol. Syst. Biol. 5, 320.

O'Brien, E.J., Monk, J.M., Palsson, B.O., 2015. Using Genome-scale Models to Predict Biological Capabilities. Cell 161, 971–987.

Opdam, S., Richelle, A., Kellman, B., Li, S., Zielinski, D.C., Lewis, N.E., 2017. A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. Cell Syst. 4, 318–329.e6.

Robinson, M.D., Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11, R25.

Simeonidis, E., Price, N.D., 2015. Genome-scale modeling for metabolic engineering. J. Ind. Microbiol. Biotechnol. 42, 327–338.

Smyth, G.K., 2005. limma: Linear Models for Microarray Data, in: Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Statistics for Biology and Health. Springer, New York, NY, pp. 397–420.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al., 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.

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

Varma, A., Palsson, B.O., 1994. Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. Nat. Biotechnol. 12, 994–998.

Vijayakumar, S., Conway, M., Lió, P., Angione, C., 2017. Seeing the wood for the trees: a forest of methods for optimization and omic-network integration in metabolic modelling. Brief. Bioinform.

Vivek-Ananth, R.P., Samal, A., 2016. Advances in the integration of transcriptional regulatory information into genome-scale metabolic models. Biosystems 147, 1–10.

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Dr. Marco Pagni (marco.pagni@sib.swiss).

METHOD DETAILS

Genome-scale metabolic network

The genome GCA_000002525.2_ASM252v1 of *Y. lipolytica* was used throughout this study. 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. For normoxia, oxygen consumption was unrestricted as in the original model and took a value of 244 $\text{mmol}\cdot\text{gDW}^{-1}\cdot\text{h}^{-1}$ 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 $\text{mmol}\cdot\text{gDW}^{-1}\cdot\text{h}^{-1}$ to simulate hypoxic conditions. Preliminary investigation showed that the model behavior did not depend too much on the exact value of this setting.

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 and Schilling, 2003) were also removed, as explained below. The finally investigated model contained 818 reactions, 469 genes, and 605 metabolites after cleaning, and is referred to as the *comprehensive* model in our analysis.

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 $\mathbf{lb}$ and $\mathbf{ub}$, 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

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

Flux Balance Analysis (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 $\mathbf{c}^{\mathrm{T}}_{\text{Biomass}}$ 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

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

and the model is said to be *viable* if $v^*_{\text{Biomass}} > 0$.

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

find min v_i and max v_i for i = 1..n

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

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

$$0 < \mathbf{c}_{\text{Biomass}}^{\text{T}} \cdot \boldsymbol{v} \leq v_{\text{Biomass}}^{*}$$

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

GO annotation

The GO annotation of *Y. lipolytica* consisted of 4629 GO terms and 4747 annotated genes (out of 6453) was retrieved from UniProt (UniProt Consortium, 2015). 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 135 GO terms (denoted *GO135* and listed in 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.

RNA-seq data

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.

QUANTIFICATION AND STATISTICAL ANALYSIS

Genome-scale metabolic network

FBA, FVA, dead-end detection, reaction and gene knock-out analyses were performed with the *sybil* R package (v2.0.4) (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 GIMME function implemented in The COBRA Toolbox v3.0 (Heirendt et al., 2017) was employed to perform analyses with the GIMME algorithm (Becker and Palsson, 2008).

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., 2013). 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 and 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*.

GO term enrichment

The analyses were done using R (v3.2.2). For comparison of enrichment methods, metaboGSE used *GO135* as input. topGO (v2.24.0) was performed independently with *classic* and *weight01* algorithms with *fisher* statistic (Alexa and 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 *GO135* 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 and Hochberg, 1995), and adjusted *p*-values < 0.05 were considered significant. For metaboGSE, single *p*-value is required to be less than 0.01 for at least one condition.

DATA AND SOFTWARE AVAILABILITY

The investigated dataset is available from the R package *metaboGSE*. The whole analysis was performed in 1.75 hours on a 64-processor Intel Xeon E5-4620 of 2.6 GHz. The parallelization is provided by the *sybil* and *parallel* packages.

SUPPLEMENTAL INFORMATION

Figure S1. Gene expression (A), fitness (B), number of genes (C), number of reactions (D), percentage of essential genes (E), and percentage of essential reactions (F) of *Y. lipolytica* sub-models from seven RNA-seq conditions, obtained by removing genes following the **expr** ranking. Only one replicate from each condition is shown. The percentages represent the numbers of essential reactions or genes in sub-models over those in the comprehensive model.

Figure S2. Comparison of genes in *Y. lipolytica* sub-models constructed with metaboGSE and GIMME from seven RNA-seq conditions. Intersection: number of common genes across conditions, Union: number of all genes across conditions, *k*: number of genes in the comprehensive model.

Figure S3. Enrichment of 38 GO terms identified via metaboGSE in *Y. lipolytica* sub-models in seven RNA-seq conditions.

Figure S4. GO terms enriched via the three methods metaboGSE, GSEA and topGO (with *classic* and *weight01* algorithms). FDR cut-off: 0.05.

Figure S5. GO terms enriched via the three methods metaboGSE, GSEA, and topGO on the ergosterol biosynthesis pathway. bp: biosynthetic process.

Figure S6. Enrichment of GO:0006696 ergosterol biosynthetic process in *Y. lipolytica* sub-models in seven RNA-seq conditions.
(A) Depletion fraction in function of number of removed genes
(B) Depletion curve: depletion fraction in function of fitness times fraction of remaining genes in the model
(C) Expression of associated genes

Table S1. Performance of different rankings ρ measured by significance level Z_ρ in all RNA-seq samples. **random**: random draw, **expr**: voom-normalized expression, **pkmExpr**: voom-normalized expression in RPKM, **relExpr*i***: relative expression $(\mathbf{expr}^i/\langle\mathbf{expr}\rangle)^{1/i}$, **revExpr**: reverse expression, **zExpr**: *z*-score. Lowest Z_ρ in bold, highest Z_ρ in italic.

Table S2. 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*. FDR from permutation test (metaboGSE), Fisher test (topGO), phenotype-based permutation test (GSEA) are cut off at 0.05. All biological process GO terms in *GO135* from Maguire *et al* with FDR are reported.