Biotechnology Journal
www.biotechnology-journal.com

Biotechnol. J. 2014, 9, 776–790

DOI 10.1002/biot.201300242

Research Article

iOD907, the first genome-scale metabolic model for the milk yeast *Kluyveromyces lactis*

Oscar Dias[1], Rui Pereira[1], Andreas K. Gombert[2], Eugénio C. Ferreira[1] and Isabel Rocha[1]

[1] CEB – Centre of Biological Engineering, Universidade do Minho, Campus de Gualtar, Braga, Portugal
[2] Faculty of Food Engineering and Bioenergy Laboratory, University of Campinas (UNICAMP), Campinas, SP, Brazil

We describe here the first genome-scale metabolic model of *Kluyveromyces lactis*, iOD907. It is partially compartmentalized (four compartments), composed of 1867 reactions and 1476 metabolites. The iOD907 model performed well when comparing the positive growth of *K. lactis* to Biolog experiments and to an online catalogue of strains that provides information on carbon sources in which *K. lactis* is able to grow. Chemostat experiments were used to adjust non-growth-associated energy requirements, and the model proved accurate when predicting the biomass, oxygen and carbon dioxide yields. When compared to published experiments, in silico knockouts accurately predicted in vivo phenotypes. The iOD907 genome-scale metabolic model complies with the MIRIAM (minimum information required for the annotation of biochemical models) standards for the annotation of enzymes, transporters, metabolites and reactions. Moreover, it contains direct links to Kyoto encyclopedia of genes and genomes (KEGG; for enzymes, metabolites and reactions) and to the Transporters Classification Database (TCDB) for transporters, allowing easy comparisons to other models. Furthermore, this model is provided in the well-established systems biology markup language (SBML) format, which means that it can be used in most metabolic engineering platforms, such as OptFlux or Cobra. The model is able to predict the behavior of *K. lactis* under different environmental conditions and genetic perturbations. Furthermore, by performing simulations and optimizations, it can be important in the design of minimal media and will allow insights on the milk yeast's metabolism, as well as identifying metabolic engineering targets for improving the production of products of interest.

Received	14 JAN 2014
Revised	07 APR 2014
Accepted	23 APR 2014
Accepted article online	28 APR 2014

Supporting information available online

Keywords: Bioinformatics · Fungi · Genome-scale metabolic model · Metabolic engineering · Systems biology

1 Introduction

Genome-scale metabolic models are now established tools utilized in a wide range of biotechnological applications, such as metabolic engineering of microbes or drug targeting [1–6]. Although a large majority of the available models are those of prokaryotes, the number of models for eukaryotic organisms has been increasing rapidly (www.optflux.org/models).

In recent years, some steps have been taken to standardize the methodology for the reconstruction of genome-scale metabolic models, for instance the publication of a detailed protocol by Thiele and Palsson [7] for the development of a standard that determines the minimum information required for the annotation of biochemical models (MIRIAM) [8]. Nevertheless, the reconstruction of the metabolic network of an organism is still a complex procedure.

The same process may, in theory, be applied for reconstructing eukaryotic and prokaryotic metabolic models [7]. Nevertheless, eukaryotic models are more demanding due to their larger knowledge base and genomes, as well as the various compartments within the cells.

Correspondence: Dr. Oscar Dias, CEB – Centre of Biological Engineering, Universidade do Minho, Campus de Gualtar, 4710-057, Braga, Portugal
E-mail: odias@deb.uminho.pt

Abbreviations: **EC**, enzyme commission; **FBA**, flux balance analysis; **GPR**, gene-protein-reaction; **KEGG**, Kyoto encyclopedia of genes and genomes; **MIRIAM**, minimum information required for the annotation of biochemical models; **P/O**, phosphorus to oxygen; **SBML**, systems biology markup language; **TCDB**, Transporters Classification Database

The reconstruction process consists of four main steps [7, 9, 10]: genome annotation, assembling the genome-scale metabolic network, conversion of the network to a genome-scale metabolic model, and finally the validation of the model. Genome annotation, in this context, is the assignment of metabolic functions, by identifying enzymes and transporters, to genes within the genome. The genome annotation allows generating gene-protein-reaction (GPR) rules, through the identification of GPR triplets (the association between the genes, the proteins encoded and the reactions promoted by such proteins). A fully annotated genome allows the assembly of a metabolic network, in which two reactions are connected if a metabolite is the substrate of one reaction and the product of the other. Besides the genome annotation results, this step usually integrates biochemical and physiological information about the specific organism available in the literature or in specialized databases. The addition of a biomass equation, constraints around the external exchange flux values, and an equation representing the depletion of adenosine triphosphate (ATP) for cellular maintenance processes, allows converting the metabolic network into a stoichiometric metabolic model. Finally, the consistency of this model can be checked by comparing simulated behavior with published experimental data. These simulations are usually performed using the flux balance analysis (FBA) formulation [11, 12].

One of the major issues when building a genome-scale metabolic model is the lack of universal identifiers for the metabolites. Unlike enzymes, which have Enzyme Commission (EC) numbers [13], and carrier proteins that are identified by Transporter Classification (TC) numbers [14], metabolites do not have an international classification standard widely accepted by the scientific community. The classification systems for metabolites that mostly resemble those for enzymes and transporters are provided by the Kyoto Encyclopedia of Genes and Genomes (KEGG) Compound database [15, 16] and MetaCyc [17]. However, only KEGG provides an application programming interface that allows retrieving this information automatically.

Tools like *merlin* ([18] and Dias, O., Rocha, M., Ferreira, E. C., and Rocha, I., Reconstructing genome-scale metabolic models with *merlin* 2.0, submitted), model SEED [19], Raven [20], MicrobesFlux [21] and others were developed specifically for model reconstruction and are becoming increasingly available. These tools are usually developed for assisting in the automation of some steps of the reconstruction process, although manual curation is always required. *merlin 2.0* is the second generation of our tool, developed for the reconstruction of genome-scale metabolic models. This user-friendly application allows performing several steps of the reconstruction process semi-automatically (Dias et al., submitted) and exporting the model in the systems biology markup language (SBML) format [22].

The yeast *Kluyveromyces lactis*, for which the complete genome sequence has been available since 2004 [23], is attracting increasing attention from molecular biologists and process engineers, and has even become a reference organism in biological research. Several aspects have contributed to this development [24–26], namely its GRAS (generally recognized as safe) status, its ability to grow on lactose as a sole carbon source, and its various industrial applications. *K. lactis* is especially useful in the dairy industry and as a host for the production of recombinant proteins [27], for which it presents an impressive secretory capacity [28, 29] and does not require methanol for efficient induction of protein production, as do methylotrophic yeasts such as *Pichia pastoris* [30]. Its distinctive petite-negative nature allows studies on mitochondrial function [31]. Moreover, the availability of various molecular tools makes it amenable to genetic manipulation [32, 33] (near the level of *Saccharomyces cerevisiae*), while its evolutionary proximity to *S. cerevisiae* allows performing comparative studies between these two species. Its regulation of carbon and energy metabolism, which contrasts with the well-studied physiology of *S. cerevisiae*, reflects its adaptation to aerobic conditions. Finally, *K. lactis* can grow on a broader diversity of substrates and is less sensitive to glucose repression than *S. cerevisiae* [34]. Like *S. cerevisiae*, *K. lactis* is an ascomycetous budding yeast that belongs to the endoascomycetales. However, whereas the *K. lactis* is an aerobic-respiring or Crabtree-negative yeast, the *S. cerevisiae* is an aerobic-fermenting or Crabtree-positive yeast [35].

The first genome-scale metabolic model for a yeast was for *S. cerevisiae* [36], for which there are currently seven metabolic reconstructions available [36–42]. Several other yeasts also have reconstructions, namely several *Pichia* strains [43–45], *Schizosaccharomyces pombe* [46], among various other fungi. However, although *K. lactis* is, along with *S. cerevisiae*, considered a prototype for modeling two distinct types of yeast, as yet there is no model for *K. lactis* [47]. A metabolic model of *K. lactis* is likely to allow comparisons that will provide relevant information on the origins of the differences between these industrially relevant yeasts and will surely permit the elucidation of interesting features of the milk yeast, as well as the identification of engineering targets for improving this organism.

Here, we present the first in silico genome-scale metabolic reconstruction of *K. lactis* with gene rules, the iOD907. This model accounts for compartmentation of reactions and the transport of metabolites across cellular membranes.

This genome-scale metabolic model complies with the MIRIAM standards for the annotation of enzymes, transporters, metabolites and reactions. Furthermore, it contains direct links to KEGG (for enzymes, metabolites and reactions) and to the Transporters Classification Database (TCDB; for transporters) allowing easy compar-

isons with other models. Finally, this model is provided in the well-established SBML format, which means that it can be used in most metabolic engineering platforms, including OptFlux [48] and COBRA [49].

2 Material and methods

2.1 Model development

Specific studies on several aspects of *K. lactis*' metabolism are lacking, including biomass composition, ATP requirements for maintenance and growth, and quantitative physiological data, such as those obtained in chemostat experiments. Thus, in addition to results from the genome annotation, other sources used to develop, improve and validate the iOD907 model were: the iMM904 *S. cerevisiae* metabolic model [41], a study by Kiers et al. in 1998 [50] on the regulation of alcoholic fermentation in *K. lactis*, Biolog Phenotype MicroArrays [51], ordered from Biolog (Hayward, CA) and publicly available data from the Centraalbureau voor Schimmelcultures – Royal Netherlands Academy of Arts and Sciences (CBS-KNAW) Fungal Biodiversity Centre webpage (http://www.cbs.knaw.nl/Collections).

The methodology used for developing the genome-scale metabolic model is depicted in Fig. 1. The main steps of this methodology are concisely described below.

2.2 Protein-reaction associations

Protein-reaction associations are available in several online databases including BRENDA [52], MetaCyc [17], or KEGG [16]. The latter was selected for this step because it provides this information automatically.

The genome annotation of *K. lactis* had been previously performed within our group [53]. Because annotations are not static and new gene functions are discovered and registered in databases every day, *merlin 2.0* was used to update this annotation. Also, the annotation of transport proteins was revised using a transporter annotation tool that was developed after the re-annotation (Dias, O., Gomes, D. G., Vilaça, P., Cardoso, J. et al., Genome-wide Semi-automated Annotation of Transporter Systems, submitted).

Using the updated annotation, the reactions associated with complete EC numbers were used to assemble the draft network. At this stage, one of the major concerns is to identify which reactions should be included in the model when KEGG associates an EC number with more

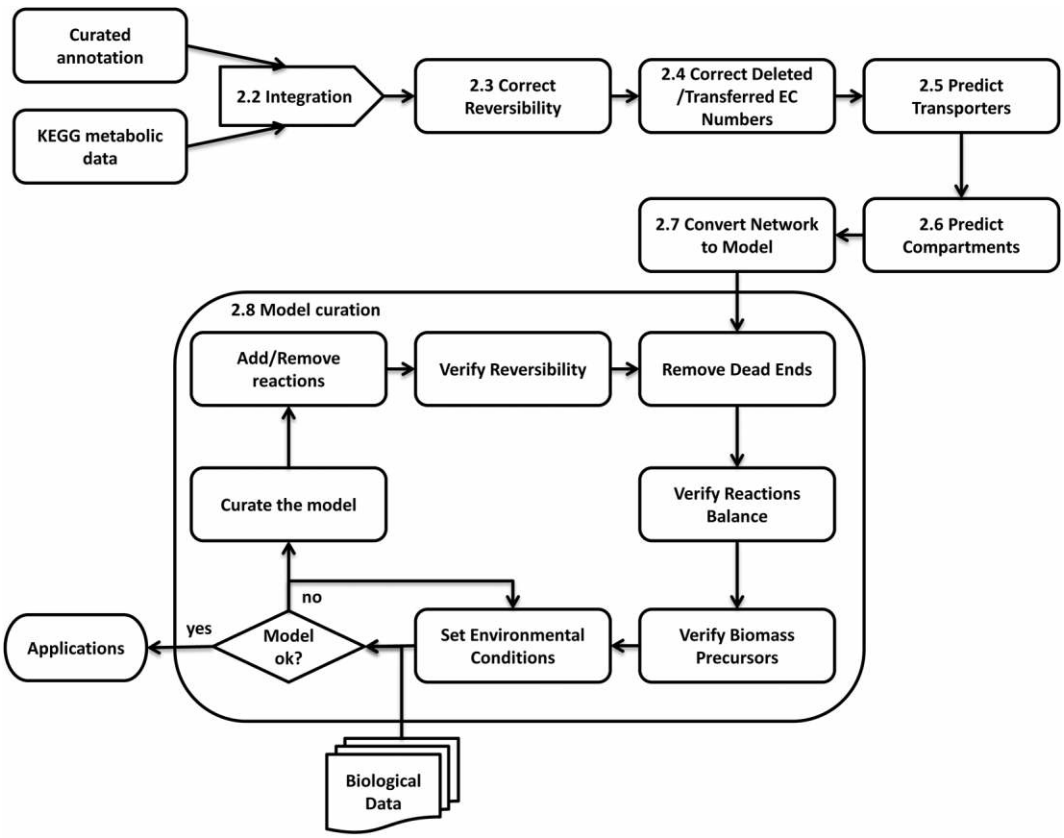

Figure 1. Methodology for the reconstruction of the *Kluyveromyces lactis* iOD907 metabolic model.

than one reaction. A conservative approach would include all reactions, but that would also create a metabolic model with many gaps and dead ends. In order to overcome that, while still having a reliable model, we used the concept of KEGG pathways. KEGG pathways are functional sets of reactions and enzymes that are connected by metabolites. The fact that an EC number is part of a pathway does not necessarily mean that all reactions associated with that EC number are also part of the same pathway. Assuming that the most relevant reactions are the ones linked to the EC numbers present in the associated pathway, in our approach, when an EC number was linked to several reactions, only those reactions present in the KEGG pathways that also included the mentioned EC number were included in the model. When the EC number only promoted a single reaction, the reaction was directly included in the model. In the same way, all reactions classified as spontaneous or non-enzymatic were also included in the model. Moreover, all reactions associated with encoded enzymes not present in any KEGG pathways were also included in the first draft of the model.

2.3 Reaction reversibility

By default, all KEGG reactions are set to be reversible. Thus, data provided in a study by Stelzer et al. [54] were used to perform an automatic initial correction of the reversibility of the reactions. These authors first retrieved the information shown in KEGG PATHWAY maps and confirmed it in BRENDA whenever possible. However, the criteria for the determination of irreversible reactions described by Ma and Zeng [55] was generally still adopted by Stelzer and co-workers when elaborating their database. Each KEGG reaction identified as irreversible in that study was automatically set to irreversible in our model.

2.4 Correct deleted/transferred EC numbers

Although the annotation was upgraded, there are no guarantees that the EC numbers available in the different databases are updated, even though the function assigned to each gene might be correct. Some EC numbers found during annotation matched KEGG records labelled as Transferred or Deleted. In these cases, a manual inspection was performed to assign roles to all metabolic genes in the model.

2.5 Transport reactions

Transport reactions were generated using genomic information together with public databases. In brief, the procedure consists of finding genes with transmembrane domains on the *K. lactis* genome using the TransMembrane prediction with Hidden Markov Models (TMHMM) approach [56]. Amino acid sequences for proteins predicted to have at least one transmembrane helix were then compared, using the Smith-Waterman [57] algorithm, to all sequences kept in the TCDB [14]. The metabolites associated to TCDB records with similarities to a given *K. lactis* gene were associated to that gene, according to the following procedure: every metabolite linked to a *K. lactis* gene was assigned a score that took into account the frequency of that metabolite among the homologous genes, as well as the taxonomy of the TCDB records associated to that metabolite and with similarities to the above-mentioned gene. The computed score, which ranged between 0 and 1, represented the likelihood of a particular metabolite being transported by a carrier encoded in that *K. lactis* gene. Hence, transport reactions for all metabolites with classifications above a given threshold were generated. The manner in which each metabolite was transported through the membrane (e.g. uniport, symport, antiport) was selected using the same process used for the sorting of metabolites. For more information on this methodology please refer to "Genome-wide Semi-automated Annotation of Transporter Systems" (Dias et al., submitted). Only transport reactions with metabolites participating in biochemical reactions were included in the model so as to avoid introducing gaps in the network.

Transport reactions from/to the exterior and across internal membranes for currency metabolites, such as H_2O, CO_2, and NH_3, which are often carried by facilitated diffusion, were added to the model with no gene association.

2.6 Compartmentation

This model accounts for four compartments: extracellular milieu, cytoplasm, mitochondrion and endoplasmic reticulum. The two internal compartments included are of utmost importance in eukaryotes since mitochondria have a major role in eukaryotic ATP synthesis, while the endoplasmic reticulum is the site where lipids, glycogen, and protein biosyntheses occur, among several other functions. The assignment of enzymes and carriers to compartments was performed using the WoLF PSORT [58] tool. Some proteins were assigned to other compartments in *K. lactis*, i.e. the nucleus, the Golgi apparatus and the peroxisome. However, the reactions catalyzed by proteins assigned to these compartments were disconnected from the network. Therefore, such enzymes were reassigned to the cytoplasm.

When discrepancies were found between predictions made by our transporter annotation tool and PSORT, preference was given to the former. Namely, non-transport reactions promoted by enzymes predicted by PSORT to be located in the plasma membrane were assigned to both the cytoplasm and the extracellular milieu, so that both possibilities would be anticipated.

On the other hand, proteins identified as carriers by the transporters annotation tool, although predicted to be localized in internal compartments by PSORT (i.e. endoplasmic reticulum or the mitochondrion), were assigned with transport reactions between the cytoplasm and the organelle. Transport reactions for carriers predicted to be localized by PSORT in the cytoplasm or the extracellular environment were discarded and assigned with transport reactions between the cytoplasm and the extracellular milieu.

2.7 Biomass formation, growth and non-growth ATP requirements

Besides the reactions from KEGG, this model includes reactions representing the formation of specific bioentities present inside the cell and reactions representing biomass formation and maintenance (non-growth) ATP requirements. These bioentities represent the average protein and the average fatty acid composition in the biomass.

Biomass formation was represented by an equation that included all components considered to be required for growth and their stoichiometries. The lack of specific studies for determining the composition of *K. lactis* was overcome by assuming that this yeast's composition was similar to the composition of *S. cerevisiae*. Hence, the biomass equation from the iMM904 *S. cerevisiae* model was used in iOD907, with a few exceptions like the proteins, nucleotides and polysaccharides contents. Table S1 (additional file 2 of the Supporting information) shows the contribution of each component that was directly extracted from iMM904.

The growth ATP requirements (also adopted from the iMM904 *S. cerevisiae* model – 59.276 moles ATP per g biomass) were introduced directly into the biomass equation.

2.7.1 Fatty acid entity

The fatty acid entity represents the average composition of the fatty acids in the cell. Again, the estimations used in the iMM904 model for *S. cerevisiae* for the weight of each fatty acid were used (Table S2, additional file 2 of the Supporting information). Fatty acids are precursors of acyl-CoA (reaction R00390), which is, in turn, a precursor of all lipids present in the biomass. The design of this entity allowed generating all lipids present in the biomass equation, i.e. phosphatidate (C00416), phosphatidylcholine (C00157), phosphatidylethanolamine (C00350), phosphatidylserine (C02737), 1-phosphatidyl-D-myo-inositol (C01194) and triacylglycerol (C00422).

2.7.2 Protein entity

Likewise, the protein entity represents the average composition of the proteins in the cell. The total protein contents were retrieved from the iMM904 (0.45 g protein per g biomass), as it was assumed that these contents are similar in yeasts. The amount of each amino acid in the protein content was estimated by calculating the percentage of each codon usage, from the translated genome sequence [7], assuming equal transcription and translation of all coding sequences, although this is not necessarily always valid [59]. Although it was not possible to experimentally validate the amino acid distribution (nor the total protein contents) in *K. lactis*, according to Santos [60], model predictions (of the specific growth rates and the flux distribution) are closer to experimental data when in silico biomass precursor coefficients are used instead of data from closely related organisms. The estimation of the amino acid contents allows focusing the model predictions on the *K. lactis* requirements encoded in the genome. The average protein composition is available in Table S3 (additional file 2 of the Supporting information).

2.7.3 Estimation of the nucleotide contents

The estimation of the nucleoside monophosphates (NMP), i.e. nucleotides, and deoxynucleoside monophosphates (dNMPs), i.e. deoxynucleotides, in the biomass can also be inferred from the genome, and were also determined using the methodology described in the protocol from Thiele and Palsson [7]. The estimation of each dNMP, shown in Table S4 (additional file 2 of the Supporting information), was performed by calculating the frequency of each nucleobase in the whole genome (including mitochondrial DNA). The same percentage of the cellular contents in DNA utilized in the iMM904 biomass equation, i.e. 0.04 g dNMP per g biomass, was used for the calculations.

The determination of the nucleotides composition, shown in Table S5 (additional file 2 of the Supporting information), was also performed according to the Thiele-Palsson protocol with one major difference: cells contain different types of RNA, which is not taken into account by the protocol, which only uses mRNA to perform these calculations. However, rRNA accounts for the majority of the RNA content in any cell; thus, in this work, three types of RNA were used: rRNA, tRNA, and mRNA, with percentages of 80, 15, and 5%, respectively [61, 62]. The same percentage of overall cellular content of RNA utilized in the iMM904 biomass equation, i.e. 0.063 g NMP per g biomass, was used for the calculations.

2.7.4 Polysaccharides

The contents of several polysaccharides present in the biomass equation, i.e. α,α-trehalose, amylose and chitin, were adapted from the iMM904 model. However, a study of the composition of the *K. lactis* cell wall [63] was used to retrieve the relative contents of 1,3-β-D-glucan and mannan in the cell. The mannan and 1,3-β-D-glucan contents in the cell are shown in Table S6 (additional file 2 of the Supporting information).

2.7.5 Phosphorus to oxygen ratio

The phosphorus to oxygen (P/O) ratio is the relationship between ATP synthesis and oxygen consumption. This quotient indicates the number of orthophosphate molecules used for ATP synthesis per atom of oxygen consumed during oxidative phosphorylation. In the absence of specific studies to characterize the P/O ratio in *K. lactis*, the same theoretical ratio (i.e. 1.5) used in the *S. cerevisiae* iMM904 metabolic model was used. The reactions contributing to this ratio were automatically generated by the transporter annotation tool. However, these reactions are generic and were updated to replicate the same P/O ratio as in the iMM904 model. The three reactions that contribute to this calculation are listed below:

Reaction T03074_C4:
1.0 *Oxygen*$_{mito}$ + 4.0 *Ferrocytochrome c*$_{mito}$ + 6.0 H^+_{mito}
⇒ 2.0 H_2O_{mito} + 4.0 *Ferricytochrome c*$_{mito}$ + 6.0 H^+_{cyto}

Reaction T03020_C4:
1.0 $Ubiquinol_{mito}$ + 2.0 $Ferricytochrome\ c_{mito}$ + 1.5 H^+_{mito}
⇒ 1.0 $Ubiquinone_{mito}$ + 2.0 $Ferrocytochrome\ c_{mito}$
+ 1.5 H^+_{cyto}

Reaction T02959_C4:
1.0 $Orthophosphate_{mito}$ + 1.0 ADP_{mito} + 3.0 H^+_{cyto}
⇒ 1.0 ATP_{mito} + 1.0 H_2O_{mito} + 3.0 H^+_{mito}

The final balance of summing these three reactions is:
3.0 $Orthophosphate_{mito}$ + 1.0 $Oxygen_{mito}$ + 3.0 ADP_{mito}
+ 2.0 $Ubiquinool_{mito}$ ⇒ 3.0 ATP_{mito} + 5.0 H_2O_{mito}
+ 2.0 $Ubiquinone_{mito}$

2.8 Model curation

The model curation protocol is described in the additional file 3 of the Supporting information. Throughout this phase, reactions were edited, and manually added to, or removed from the model. The manual inspection of the fluxes associated to reactions involved in the formation of the biomass precursors exposed some gaps in the network. Whenever a gap was found in the model, reactions were sought to fill that gap. The KEGG pathways and the iMM904 model were used as standards. If the model lacked a reaction in a KEGG pathway, this gap was analyzed to search for additional GPR evidence and the reaction was added to the network. If the gap was outside a KEGG pathway, gap filling reactions identified within the iMM904 model were sought in KEGG. If the reaction was not available in KEGG, it was manually created and added to the network. In either case, the model was improved with the new reaction and the gap was filled.

This process was repeated several times, according to Fig. 1, until the in silico results replicated the in vivo data. The model was tested by simulating growth using the environmental conditions presented in Table S7 (additional file 2 of the Supporting information). This methodology was implemented using *merlin 2.0* for the reconstruction process and OptFlux 3.0 [48] for the validation of the model. All predictions were performed using the IBM® CPLEX solver.

2.9 GPR associations

The relationship between genes, proteins and reactions in the cases of non-one-to-one associations was automatically retrieved, by *merlin,* from pathway modules and complex modules provided in the KEGG BRITE database [64]. The Boolean rules determined whether the genes are associated to a protein and reaction by an AND rule (protein complexes) or an OR rule (isoenzymes). Nevertheless, some rules were determined by the authors' previous knowledge, such as the determination of rules for known protein complexes, like the PFK (phosphofructokinase) complex, which could not be determined by the GPR tool.

3 Results and discussion

3.1 Model characteristics and validation

Comparison of the automatic annotation, performed with the default *merlin 2.0* parameters, with the previous one [53], resulted in the updating of the annotation of 45 genes and the addition of 22 new metabolic genes and 1 non-catalytic essential subunit added by GPR rules (Table S8, additional file 2 of the Supporting information). The new metabolic genes, not present in the previous annotation, included 8 genes encoding enzymes and 14 genes encoding carriers.

From the 1788 metabolic genes provided by the updated annotation (1759 from the previous annotation + 6 previously discarded genes + 23 new metabolic genes), only 906 were used in the final version of the metabolic model. The remaining 881 genes are available for subsequent development of an extended version of the model in *merlin 2.0*, but were excluded for several reasons:

- Approximately one quarter (204) of these genes encoded enzymes exclusively identified with partial EC numbers, and thus were not integrated in the model.
- 82 exclusively encoded transporters. However, the transport reaction generation tool did not assign any reaction to 23 of these genes, so it was not possible to include them in the model. The remaining 59 genes are connected to transport reactions in *merlin 2.0*, but the corresponding metabolites were not present in the metabolic model.
- The remaining 595 genes encoded proteins promoting reactions available in *merlin 2.0*, but were not included in this version of the model either because the metabolites are disconnected from the main network or due to a decision taken during the manual curation

of the model. For instance, genes BDH1_KLULA (KLLA0F00582g) and BDH2_KLULA (KLLA0F00594g) encode enzymes 1.1.1.303 and 1.1.1.4, which promote reactions R02855 and R02946, respectively. Although initially present in the model, these genes and reactions were removed as they were disconnected from the remaining network.

Nevertheless, the final version of the iOD907 model included 906 metabolic genes + 1 gene encoding a non-metabolic essential subunit. These genes are associated with 1867 reactions (1107 internal and 760 transport) involving 1476 species in four different compartments (the same metabolite in different compartments is considered a distinct species).

As shown in Table S9 (additional file 2 of the Supporting information), only a small number of the genes in this model (154) had their annotation confirmed by UniProt, at the time of the development of the model. The annotation status of the remaining 753 genes was "unreviewed" and often the automatic annotations available in UniProt did not assign them any function. Most of the genes with reviewed annotations are associated with enzymes present in the central carbon metabolism.

Although it might seem that the number of genes in this model is similar to the number of genes in the *S. cerevisiae* model, i.e. the widely used iMM904, it should be kept in mind that baker's yeast has experienced whole genome duplication [65] and many reactions in this model might be connected to paralogous genes. Thus, proportionally, the number of reactions is much higher in the iOD907 (1043 reactions associated with 904 genes in iMM904, and 1785 reactions associated with 907 genes in iOD907).

Finally, the effective in silico P/O ratio, calculated according to Famili et al. [66], for *K. lactis* is 1.04. The difference between the theoretical and effective P/O ratio can be explained by the transmembrane proton gradient needed to carry metabolites across the membranes, such as in the symport of proton-coupled pyruvate [66], as demonstrated with simulations performed on the iOD907 model.

3.2 Oxygen availability

The iMM904 *S. cerevisiae* model can simulate anaerobic growth when the environmental conditions are supplemented with sterols (ergosterol and zymosterol) and unsaturated fatty acids (C16-C18), because the biosynthesis of these metabolites requires oxygen, and when heme (only mandatory for oxidative phosphorylation) is removed from the biomass equation, resembling the in vivo behavior. However, although in silico anaerobic growth is possible with the iOD907, *K. lactis* is not capable of surviving under these conditions, even if supplemented with sterols and unsaturated fatty acids. According to Snoek and Steensma [67], one of the reasons for this could be the lack of genes involved in sterol uptake. However, the transporters annotation tool found several *K. lactis* genes with homologies to the genes of *S. cerevisiae* known to be related to such function, as shown in Table S10 (additional file 2 of the Supporting information). For instance, the gene KLLA0D18601g was found to be homologous to *S. cerevisiae ARV1* gene, known to be required for sterol uptake and for growth during anaerobiosis. Another reason proposed by the authors for this behavior is the absence of transcription factors involved in sterol uptake. However, this was not corroborated in our work. During the re-annotation of the genome it was noticed that the KLLA0A04169g gene is a functional homologue of the *S. cerevisiae UPC2* gene, known to be implicated in the activation of anaerobic genes involved in sterol uptake and regulation of sterol biosynthesis. Therefore, the absence of anaerobic growth in *K. lactis* does not seem to be related to any metabolic deficiency nor correlated with the regulation of sterol uptake, and may rather be associated with several other factors, e.g. other regulatory phenomena, as also remarked on by the authors [67]. Since the iOD907 does not include any regulatory mechanisms, this would justify the fact that the predictions obtained do not match the real behavior.

Although simulations performed under oxygen-limiting conditions (1 mmol O_2.gDw$^{-1} \cdot$ h^{-1}) predict the production of ethanol, it is generally accepted that under such conditions *K. lactis* starts increasing glucose metabolism, accumulating both ethanol and glycerol. However, it is inaccurate to evaluate by-product formation by only inspecting FBA results from one simulation. In fact, as different combinations of by-products often offer equivalent stoichiometric results, it is necessary to perform flux variability analysis (FVA) [68] to interpret predictions of by-product formation.

Therefore, the minimum and maximum in silico yields of the most common fermentation by-products (acetate, ethanol, glycerol, pyruvate and succinate) were assessed using uptake fluxes (environmental conditions in OptFlux) corresponding to limited glucose availability and low oxygen concentration (Env 1: $q_{O2} = 1.2$ mmol.g$^{-1} \cdot$ h^{-1}, $q_{glucose} = 2.49$ mmol.g$^{-1} \cdot$ h^{-1}; Env 2: $q_{O2} = 1.7$ mmol.g$^{-1} \cdot$ h^{-1}, $q_{glucose} = 2.045$ mmol.g$^{-1} \cdot$ h^{-1}, as described in Table S7, additional file 2 of the Supporting information), according to a study on the dependence of baker's yeast on oxygen for energy generation [69]. The minimum biomass flux was set to the yield (b_m) obtained when maximizing biomass in each environmental condition for both the *K. lactis* and the *S. cerevisiae* models. Through evaluation of these yields it was possible to determine whether there were any significant differences, stoichiometrically, in the production of a given by-product compared to another. The results of this evaluation are shown in Fig. 2.

As shown in Fig. 2, the only mandatory fermentation by-product is ethanol. The in silico yield for this metabo-

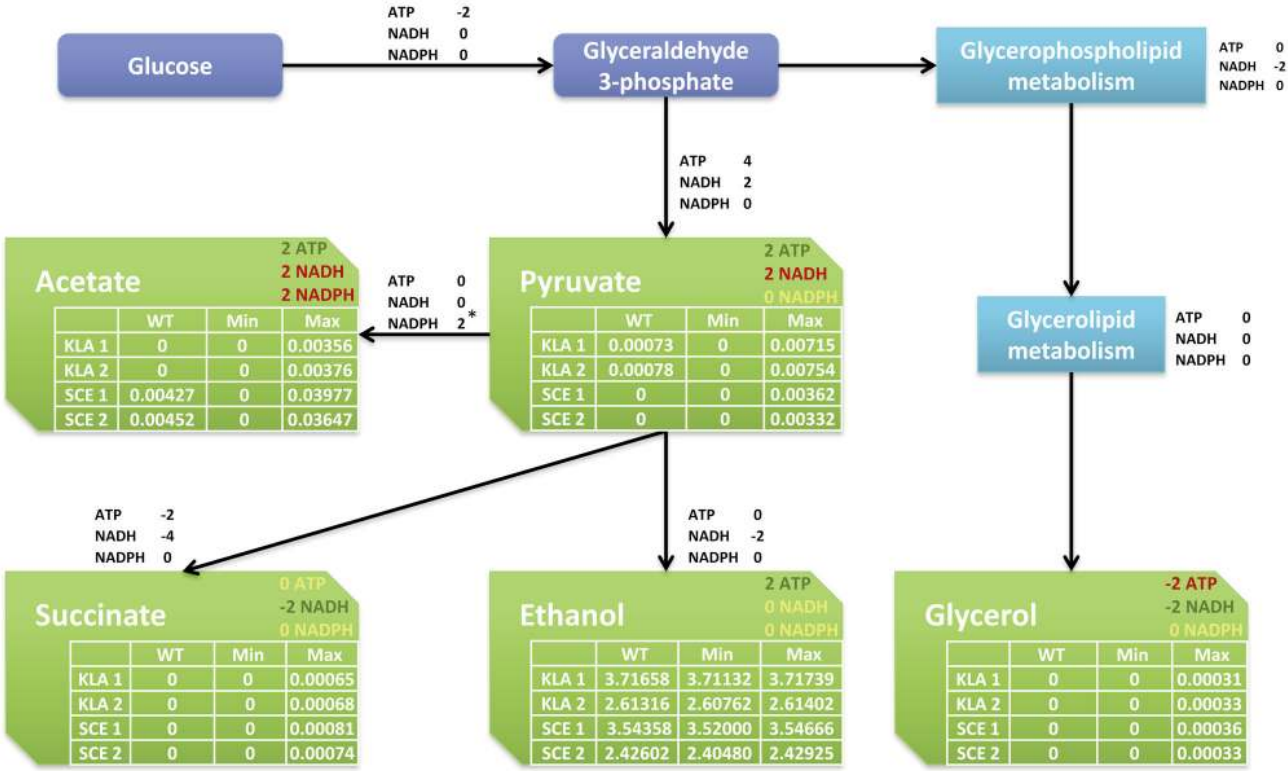

Figure 2. Analysis of the energy balances involved in the formation of several by-products (acetate, ethanol, glycerol, pyruvate and succinate). *K. lactis* wild-type simulation 1 (KLA1) and *S. cerevisiae* wild-type simulation 1 (SCE1) were obtained using ENV1 environmental conditions. Likewise, *K. lactis* wild-type simulation 2 (KLA2) and *S. cerevisiae* wild-type simulation 2 (SCE2) use ENV2. Env 1: $q_{O2} = 1.2$ mmol.g$^{-1}\cdot$h^{-1}, $q_{glucose} = 2.49$ mmol.g$^{-1}\cdot$h^{-1}; Env 2: $q_{O2} = 1.7$ mmol.g$^{-1}\cdot$h^{-1}, $q_{glucose} = 2.045$ mmol.g$^{-1}\cdot$h^{-1}. The energetic theoretical yields (ATP, NADH and NADPH) were calculated assuming one molecule of glucose. The formation of all by-products occurs via glycolysis. Glycerol is produced from glyceraldehyde 3-phosphate, thus it does not account for the 4 ATP molecules generated when glyceraldehyde 3-phosphate is converted to pyruvate, yielding 2 ATP molecules. On the other hand, the formation of pyruvate consumes 2 NAD^+ molecules, while the glycerophospholipid metabolism, which is an intermediate in the production the production of glycerol, generates 2 NAD^+ molecules. The conversion of pyruvate to ethanol is coupled to the consumption of 2 NADH molecules, thus the excretion of ethanol is very favorable under oxygen limitation. Similarly, the conversion of pyruvate to succinate consumes 4 NADH molecules. However, 2 ATP molecules are also consumed for the generation of this by-product. The conversion of pyruvate to acetate consumes 2 $NADP^+$ and 2 NAD^+ molecules. Therefore, although the production of acetate yields 2 ATP molecules, the overall yields are unfavorable as there is no net co-factor regeneration. (* – assuming an NADPH-dependent acetaldehyde dehydrogenase, similar to *S. cerevisiae*' ALD6 [70], in *K. lactis*).

lite is very robust, since the maximum and minimum yields vary by less than 1% from the original FBA simulation for both yeasts and environmental conditions. On the other hand, the production of all other metabolites is optional, since a value of zero for each metabolite still provides a viable phenotype.

The energy balance clearly favors the formation of ethanol, since the excretion of this metabolite is associated with a theoretical yield of 2 ATP molecules per mole of glucose, and neutral yields for both NADH and NADPH. When growing under oxygen-limiting conditions, the utilization of the reduced form of these coenzymes for generating energy is compromised due to the lack of oxygen. In addition, the oxidized form of NAD^+ is required for instance to oxidize the carbon source. Therefore, the regeneration of NAD^+ and $NADP^+$ under these conditions is essential to the cell.

The excretion of pyruvate or acetate also provides 2 ATP per mole of glucose for growth and cellular maintenance. However, the metabolic pathway for the production of both metabolites requires the reduction of 2 NAD^+ molecules and, in the case of acetate, an additional amount of 2 $NADP^+$ molecules.

Succinate, although yielding 2 NAD^+ molecules, has neutral theoretical ATP and NADPH yields. Cells have high growth and non-growth ATP requirements [9, 71], so the formation of this by-product is not as beneficial for the cell as ethanol, unless extra NADH needs to be recycled.

Similarly, glycerol production yields 2 NAD^+ and has a neutral theoretical NADPH yield. However, the formation of glycerol has a yield of –2 ATP molecules, which is very unfavorable for the cell in oxygen-limiting conditions in which the ATP availability is tightly controlled.

Furthermore, as shown in the Supporting information, Fig. S1 (additional file 3), formate is a mandatory by-product of the biosynthesis of several metabolites necessary for cellular growth, i.e. steroids and coenzymes (riboflavin and FAD^+). Usually, when oxygen is not limiting, this metabolite is oxidized to CO_2 by an acceptor, which, in turn, is oxidized directly or indirectly by O_2. However, these environmental conditions make the cell redirect the oxygen to essential reactions, such as the biosynthesis of heme and lipids. Therefore, its formation is mandatory.

3.3 Gap filling

The 84 reactions used to fill gaps in the iOD907 model are listed in Table S11 (additional file 2 of the Supporting information). This list includes 50 reactions (27 enzymatic + 12 spontaneous + 11 non-enzymatic) added to the model without genomic evidence but which are present in the KEGG. The added enzymatic reactions often had incomplete EC numbers associated with them, impairing the direct inference of reactions from the annotation results. The list of gap filling reactions also includes 32 reactions (27 transport + 5 other) added to the model without genomic evidence and which are not available in the KEGG. These transport reactions were not predicted by the transporter annotation tool and were usually for currency metabolites (i.e. metabolites with hundreds of connections in the model, e.g. water, oxygen, ammonia) and for the transport of specific metabolites to uncommon compartments, like NADPH and ergosterol to the endoplasmic reticulum. One of the explanations for this is that the methodology for the generation of transport reactions is very stringent, because the tool has to meet several criteria to generate such reactions. Lastly, the list of gap filling reactions includes 2 reactions with GPRs, not available in the KEGG, adopted from the iMM904 model. In this case, KEGG reactions were modified to resemble the imm904 reactions. For instance, R_SUCD3_u6m from the iMM904 model is not available in the KEGG, yet it resembles reaction R02164 from the KEGG that uses fumarate as an electron acceptor. This cofactor was replaced by FAD^+ to match the iMM904 reaction, as shown in reaction R02164_FAD_C4. These reactions were used either to eliminate critical gaps in the network or because the reactions can take place without the intermediation of a catalyst.

3.4 Carbon sources

The growth on different carbon sources with the iOD907 model was assessed by comparing in silico predictions to in vivo experiments obtained using Biolog Phenotype MicroArrays and information obtained from the CBS-KNAW on *K. lactis* CBS 2359, as shown in Table S12 (additional file 2 of the Supporting information). The combination of all the information from both data sources resulted in a set of 199 metabolites in which the growth of *K. lactis* was tested using at least one of the information sources. As shown in Table 1, only 21 carbon sources tested in vivo by both Biolog and CBS-KNAW could be tested in silico. Thus, the following discussion will focus on these carbon sources. An extensive analysis of all carbon sources tested in every dataset is available in the additional file 3 of the Supporting information.

Table 1. *Kluyveromyces lactis* growth assessment for the carbon sources tested in all data sets (in silico, Biolog assays and CBS-KNAW). Growth (+) and lack of growth (–) were verified.

Carbon source	KEGG ID	Biolog	CBS-KNAW	in silico
Sucrose	C00089	+	+	+
D-Xylose	C00181	+	+	+
Maltose	C00208	+	+	+
D-Lactose	C00243	+	+	+
D-Glucose	C00267	+	+	+
α,α-Trehalose	C01083	+	+	+
Succinate	C00042	–	+	+
Glycerol	C00116	–	+	+
D-Galactose	C00124	–	+	+
Citrate	C00158	–	+	+
L-Lactate	C00186	–	+	+
Xylitol	C00379	–	+	+
Raffinose	C00492	–	+	+
D-Sorbitol	C00794	–	+	+
D-Ribose	C00121	+	–	+
D-Glucosamine	C00329	+	–	+
D-Gluconate	C00257	–	–	+
L-Lysine	C00047	–	+	–
myo-Inositol	C00137	–	–	–
D-Glucuronate	C00191	–	–	–
Melibiose	C05402	–	–	–

As shown in both tables mentioned above, consensus between the two data sources and the in silico predictions (iOD907) was only attained for 9 carbon sources (establishing growth in 6 and not growing in 3) out of a total of 21. Moreover, as show in Table 1, the iOD907 positive-growth predictions agreed with CBS-KNAW for 8 additional carbon sources in which no growth was indicated with the Biolog experiments (surprisingly, this list includes one of the carbon sources in which *K. lactis* is known to thrive, D-galactose, unlike other yeasts such as *S. cerevisiae*). On the other hand, as show in Table 1, the iOD907 model matched growth with the Biolog experiments for 2 other carbon sources in which CBK-KNAW did not establish growth. In addition, there is 1 carbon source, which is reported by CBS-KNAW to be a viable carbon source, for which growth could not be identified in either the in silico strain or the Biolog experiments (L-lysine). Again, surprisingly, the in silico strain was able to use D-gluconate as sole carbon source for growth, despite the fact that both Biolog and CBS suggest no growth in in vivo experiments.

It should be noted that the ability to grow in a particular carbon source is often dependent on not only the enzymatic and transport capabilities of the organism, but also the presence or absence of other medium components. A possible explanation for the discrepancies found between the experiments is the different setups of each test, i.e. the absence or presence of other nutrients.

In conclusion, the compatibility of the results from the in vivo growth tests performed by Biolog and CBS-KNAW, and the in silico simulations are fairly positive since it encompasses 9 out of 21 possible carbon sources. Within these carbon sources, there was inconsistency between the two in vivo data sources and the model predictions in one case, an agreement between in silico and Biolog in three cases and between in silico and CBS-KNAW in eight cases. All the carbon sources for which the results of the in vivo tests were inconsistent should thus be double checked to confirm or refute the in silico prediction.

3.5 Maintenance ATP fitting

The depletion of ATP by processes not directly associated with growth, like futile cycles or turnover of molecules, was represented in the model by an equation that forces ATP consumption via a specific flux. The boundaries of this flux were inferred by fitting the in silico predictions of the model to experimental in vivo data from Kiers et al. [50].

The model was used to predict growth, oxygen consumption and carbon dioxide production yields using the same environmental conditions utilized in that work, and limiting the carbon source (glucose) availability to the actual glucose uptake rate, using different maintenance ATP flux values (1.0–5.0 $mmol.h^{-1} \cdot g^{-1}$). Table S13 (additional file 2 of the Supporting information) lists the results for the simulations performed with OptFlux.

The predictions of the models for each maintenance ATP value were fitted to the in vivo data. The linear regression slopes and y-intercept values for each maintenance ATP regression are depicted in the Supporting information, Fig. S2 (additional file 3).

The analysis of the above-mentioned figure shows that the ATP used for maintenance should be set to 2 $mmol.gDW^{-1} \cdot h^{-1}$, as this value provides the best overall fitting to the in vivo data on all the predictions.

3.6 Knockout analysis

The results of several gene deletions performed in in vivo experiments were collected and the corresponding phenotypes compared to the in silico predictions of this model to assess the model reliability. The result of this comparison is shown in Table S14 (additional file 2 of the Supporting information).

As shown, over 90% of the model predictions are in accordance with the simulation results, thus confirming the accuracy of the model. The model correctly predicted 25 true positives and 15 true negatives. On the other hand, the in silico simulations predicted 2 viable mutants for which there was no growth in vivo and only 1 false negative.

One of the main differences between *K. lactis* and *S. cerevisiae* is the former's viable *RAG2* mutant phenotype [53, 72]. The viability if this mutant is correctly predicted by the iOD907 model. The simultaneous deletion of *RAG2* and *TAL1*, although not impairing growth in non-fermentable carbon sources in vivo [73], is lethal for *K. lactis* in silico (false negative). As shown in Table S14 (additional file 2 of the Supporting information), the deletion of these genes separately is not critical, because the other gene can be used to bypass the deletion; however, the deletion of both genes impairs glycolysis and gluconeogenesis. Such discrepancy might be associated to the strain used in the in vivo study (HK5-2B), which was not the one used to develop the in silico model (NRRL-Y1140/CBS 2359/ATCC 8585).

Another distinction from *S. cerevisiae* is that the *PDC1* (pyruvate decarboxylase) null mutation in *K. lactis,* for growth on glucose, attains the same yield as the wild type [74], which is also predicted in this model. In addition, this mutant did not accumulate ethanol under oxygen-limited conditions (oxygen flux limited to 1 $mmol.gDw^{-1} \cdot h^{-1}$), accumulating pyruvate instead.

Surprisingly, the in vivo deletion of each of the phosphofructokinase (PFK) subunits by themselves did not impair *K. lactis* growth on fermentable carbon sources. Jacoby and colleagues [73, 75] claimed that "This could be caused by a residual PFK activity conferred by the remaining subunit in vivo that escapes detection by in vitro enzymatic determinations". Indeed, the two PFK subunit sequences have a similarity of over 40%, which means that the deletion of one subunit may be partially bypassed by the remaining subunit. However, in this model, the deletion of either of the subunits will remove the reactions associated with this complex, since there is a gene rule that associates the presence of both genes to that enzymatic activity. Nevertheless, the non-growth in fermentable carbon sources for the *PFK1* mutant in conjunction with the *TAL1* knockout is correctly predicted by the iOD907.

Similarly, the separate deletion of the *ARG8* and *LYS2* genes generated auxotrophic mutants on arginine and lysine, respectively. This phenotype was also observed in silico. The in silico deletion of the *ICL1* gene produced mutants that did not grow on ethanol, which is in accordance with the in vivo data. However, the in vivo data deletion of *FBP1* does not agrees with the in silico prediction. Just as in glycolysis the deletion of the *RAG2* gene or the phosphofructokinase complex is bypassed by the pentose phosphate pathway (PPP), in the gluconeogenesis the deletion of *FBP1* may be bypassed by the inverse route in the PPP. In silico, the lack of data on the reversibil-

ity of the reactions in the PPP can be compensated by an alternative route for the generation of glucose from ethanol when the fructose-1,6-bisphosphatase is deleted.

As the iOD907 is a stoichiometric model, the predictions of the decreased growth rate when the dominant acetyl-coenzyme A synthetase copy (*ACS1* gene) is deleted or the less-decreased growth rate provided by the *ACS2* knockout, could not be verified, although the viability was confirmed for these single deletions. The double mutant lethal phenotype was verified in silico, as the model did not predict growth on glucose, acetate or ethanol.

The in silico growth rate (on glucose) of the *PDA1* knockout mutant was decreased when compared to the wild type, which, although not the fourfold reduction found by Zeeman et al. [76], is in accordance with the in vivo experiments. The deletion of the only gene encoding an invertase in *K. lactis* (*INV1*) did not impair growth on glucose, but it was lethal for growth on raffinose. Although in the *K. lactis* annotation this is the only enzyme able to hydrolyze polysaccharides, the authors of the study [34] only report defective growth on raffinose when the gene encoding this enzyme is knocked out.

The knockout of the *TPS1* gene prevents *K. lactis* from growing on glucose or fructose, both in vivo and in silico. However, this mutant is viable when using α,α-trehalose as carbon source. Again, the stoichiometric nature of the model cannot predict the reduction of the growth rate in this mutant proposed by the in vivo experiments, as this reduction may arise from a decreased affinity for α,α-trehalose uptake, among several other factors.

In contrast to *S. cerevisiae*, for which the deletion of the *TPI1* gene is lethal, the phenotype of the deletion of this gene in *K. lactis* is viable. In the latter case, this mutation increases the glycerol yield under oxygen-limited conditions (oxygen flux limited to 1 $mmol \cdot gDw^{-1} \cdot h^{-1}$), which can be verified in silico. The in silico strain cannot predict the formation of glycerol in the presence of this gene, because all the glycerone phosphate produced by the fructose-bisphosphate aldolase is redirected to glycolysis by this enzyme, instead of generating glycerol. The viability of this mutant is related to the bypassing of the glycolytic flux through the PPP, as previously clarified in [53].

3.7 *K. lactis* versus other yeasts

As shown in Table 2, *K. lactis* has very distinctive characteristics when compared to other yeasts for which genome-scale models are available. *K. lactis* is an obligate aerobic respiring yeast, as are *S. pombe*, *Scheffersomyces stipitis* (*Pichia stipitis*) and *P. pastoris*. However, this particularity could not be mimicked by the model because, as shown previously, when deprived of oxygen and supplemented with certain metabolites, anaerobic growth can be achieved in silico. Therefore, its obligate aerobic nature has to be associated by regulatory phenomena.

The Crabtree-negative nature of *K. lactis* is only shared with the two *Pichia* species, since *S. pombe*, like the facultative anaerobe *S. cerevisiae*, is Crabtree positive. These characteristics cannot be confirmed in this model, because the internal fluxes are unrestricted. The restriction of the internal fluxes implies fluxomics studies, which would allow setting maximum flux values in specific reactions.

K. lactis is the only yeast that can metabolize xylose, lactose and galactose, a fact that is confirmed by iOD907. Although *P. stipitis* is also supposed to grow on these three carbon sources, growth on lactose was reported as variable. All other yeasts discussed above are reported as not growing on these sugars. We could not find knockout mutant studies for the *Pichia* species. For the yeasts with available mutant phenotype assays, *K. lactis* is the only one that has a viable *RAG2* mutant that is also replicated in the genome-scale model. The knockout of the *PDC1* and *TPI1* genes generated viable phenotypes in *K. lactis*

Table 2. Comparison of particular metabolic characteristics of yeasts with currently available metabolic models[a)]

Property	iOD907	KLA	SCE	SPO	PST	PPA
Crabtree effect	–	Negative	Positive	Positive [77]	Negative [77]	Negative [78]
Full anaerobic growth	Yes	No	Yes	No [79]	No [80]	No [81]
Alternate carbon sources						
Xylose	Yes	Yes	No	No [46]	Yes [43]	V [82] / No[b)]
Lactose	Yes	Yes	No	No [46]	V [82]	No [82][b)]
Galactose	Yes	Yes	No	No [46]	Yes [82]	No [82][b)]
RAG2	Yes	Yes	No	No [46]	N/A	N/A
Viable mutants						
PDC1	Yes	Yes	No	Yes [46][c)]	N/A	N/A
TPI1	Yes	Yes	No	Yes [46]	N/A	N/A

a) KLA, *Kluyveromyces lactis*, SCE, *Saccharomyces cerevisiae*, SPO, *Schizosaccharomyces pombe*, PST, *Scheffersomyces stipitis (Pichia stipitis)*, PPA, *Pichia pastoris*, V, variable.
b) http://www.cbs.knaw.nl/collections/BioloMICS.aspx?Table=Yeasts%20species&Name=Pichia%20pastoris&ExactMatch=T
c) Has paralogues.

and *S. pombe*. All viable phenotypes in *K. lactis* were predicted by the iOD907 model.

3.8 Other properties

The presence of folate in the environmental conditions is not mandatory; however, its exclusion from the growth medium has a side effect: the mandatory production of glycoaldehyde. In the folate biosynthesis pathway the reaction catalyzed by the dihydroneopterin aldolase (4.1.2.25), which produces 2-amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine that is needed to produce folate, generates glycoaldehyde. However, this metabolite is not reused in the network, and thus has to be excreted by the cell. Although this is also observed in the *S. cerevisiae* model, to the best of our knowledge there is still not an experimental evaluation of this phenomenon.

The net conversions for the three environmental conditions utilized in this work are available in Table S15 (additional file 2 of the Supporting information). Likewise, the reactions and respective fluxes for these environmental conditions are described in Table S16 (additional file 2 of the Supporting information).

4 Concluding remarks

This model was developed semi-automatically using *merlin 2.0* and a previous genome-wide re-annotation of the *K. lactis* genome, allowing a fast reconstruction (in a couple of months).

The excretion of by-products during growth under hypoxic conditions in iOD907 was assessed. It was shown that the only by-product with robust fluxes in this model was ethanol. The production of all other metabolites is facultative, as it not appear to lend any significant advantages to this organism.

The iOD907 in silico model performed well when comparing the positive growth of *K. lactis* to ordered Biolog experiments and to an online catalogue of strains (CBS-KNAW) that also provides information on growth-associated carbon sources. The model was in agreement with both data sources for 9 carbon sources. Consensus between Biolog, CBS-KNAW and in silico simulations could not be obtained for 12 other carbon sources, meaning that there is room for improvement of this model, although there are also errors arising from the Biolog growth assays, e.g. lack of growth on galactose.

The iOD907 model was able to predict phenotypes for more than 90% of the knockouts from several experiments published over the last three decades. Moreover, it provides reasonable results for quantitative simulations of chemostat experiments, as shown in the previous section. Those knockouts include experiments that mark the difference between *K. lactis* and *S. cerevisiae*, such as the viable *PDC1* and *TPI* mutants, which were predicted by iOD907. These results clearly demonstrate that, despite having similarities with the baker's yeast model, the *K. lactis* model has its peculiarities and clear distinctions from the *S. cerevisiae* models.

This model will allow insights on the metabolism of milk yeast, as well as identifying engineering targets for improving the production of products of interest by performing in silico simulations. It is freely available on the following website: www.*merlin*-sysbio.org/files/iOD907.xml (additional file 1 of the Supporting information).

The authors thank strategic Project PEst-OE/EQB/LA0023/2013 and project "BioInd – Biotechnology and Bioengineering for improved Industrial and Agro-Food processes, REF. NORTE-07-0124-FEDER-000028" co-funded by the Programa Operacional Regional do Norte (ON.2 – O Novo Norte), QREN, FEDER. The authors would also like to acknowledge Steve Sheridan for proof reading this manuscript.

The authors declare no financial or commercial conflict of interest.

5 References

[1] Bro, C., Regenberg, B., Förster, J., Nielsen, J., In silico aided metabolic engineering of *Saccharomyces cerevisiae* for improved bioethanol production. *Metab. Eng.* 2006, *8*, 102–111.
[2] Brochado, A. R., Matos, C., Møller, B. L., Hansen, J. et al., Improved vanillin production in baker's yeast through in silico design. *Microb. Cell Fact.* 2010, *9*, 84.
[3] Lee, S. J., Lee, D.-Y., Kim, T. Y., Kim, B. H. et al., Metabolic engineering of *Escherichia coli* for enhanced production of succinic acid, based on genome comparison and in silico gene knockout simulation. *Appl. Environ. Microbiol.* 2005, *71*, 7880–7887.
[4] Asadollahi, M. A., Maury, J., Patil, K. R., Schalk, M. et al., Enhancing sesquiterpene production in *Saccharomyces cerevisiae* through in silico driven metabolic engineering. *Metab. Eng.* 2009, *11*, 328–334.
[5] Terstappen, G. C., Reggiani, A., In silico research in drug discovery. *Trends Pharmacol. Sci.* 2001, *22*, 23–26.
[6] Kim, T. Y., Kim, H. U., Lee, S. Y., Metabolite-centric approaches for the discovery of antibacterials using genome-scale metabolic networks. *Metab. Eng.* 2010, *12*, 105–111.
[7] Thiele, I., Palsson, B. Ø., A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* 2010, *5*, 93–121.
[8] Le Novère, N., Finney, A., Hucka, M., Bhalla, U. S. et al., Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.* 2005, *23*, 1509–1515.
[9] Rocha, I., Förster, J., Nielsen, J., Design and application of genome-scale reconstructed metabolic models. *Methods Mol. Biol.* 2008, *416*, 409–431.
[10] Francke, C., Siezen, R. J., Teusink, B., Reconstructing the metabolic network of a bacterium from its genome. *Trends Microbiol.* 2005, *13*, 550–558.
[11] Papoutsakis, E. T., Equations and calculations for fermentations of butyric acid bacteria. *Biotechnol. Bioeng.* 1984, *26*, 174–187.

[12] Savinell, J. M., Palsson, B. O., Optimal selection of metabolic fluxes for in vivo measurement. I. Development of mathematical methods. *J. Theor. Biol.* 1992, *155*, 201–214.
[13] Barrett, A. J., Canter, C. R., Liebecq, C., Moss, G. P. et al., *Enzyme Nomenclature*. San Diego: Academic Press, 1992, p. 862.
[14] Saier, M. H., Tran, C. V, Barabote, R. D., TCDB: The Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res.* 2006, *34*, D181–186.
[15] Goto, S., Okuno, Y., Hattori, M., Nishioka, T. et al., LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* 2002, *30*, 402–404.
[16] Kanehisa, M., Goto, S., KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000, *28*, 27–30.
[17] Caspi, R., Altman, T., Dreher, K., Fulcher, C. A. et al., The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 2012, *40*, D742–753.
[18] Dias, O., Rocha, M., Ferreira, E. C., Rocha, I., *Merlin*: Metabolic models reconstruction using genome-scale information. In: Banga, J. R., Bogaerts, P., Van Impe, J., Dochain, D., Smets, I. (Eds.), *Proceedings of the 11th International Symposium on Computer Applications in Biotechnology (CAB 2010)*, 2010, pp. 120–125.
[19] Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M. et al., High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* 2010, *28*, 977–982.
[20] Agren, R., Liu, L., Shoaie, S., Vongsangnak, W. et al., The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. *PLoS Comput. Biol.* 2013, *9*, e1002980.
[21] Feng, X., Xu, Y., Chen, Y., Tang, Y. J., MicrobesFlux: A web platform for drafting metabolic models from the KEGG database. *BMC Syst. Biol.* 2012, *6*, 94.
[22] Hucka, M., Finney, A., Sauro, H. M., Bolouri, H. et al., The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 2003, *19*, 524–531.
[23] Dujon, B., Sherman, D. J., Fischer, G., Durrens, P. et al., Genome evolution in yeasts, *Nature* 2004, *430*, 35–44.
[24] Gonzales-Siso, M. I., Freire-Picos, M. A., Ramil, E., Gonzalez-Dominguez, M. et al., Respirofermentative metabolism in *Kluyveromyces lactis*: Insights and perspectives, *Enzyme Microb. Technol.* 2000, *26*, 699–705.
[25] Schaffrath, R., Breunig, K. D., Genetics and molecular physiology of the yeast *Kluyveromyces lactis*. *Fungal Genet. Biol. FG B*, 2000, *30*, 173–190.
[26] Micolonghi, C., Corsi, E., Conte, R., Bianchi, M. M., Heterologous products from the yeast *Kluyveromyces lactis*: Exploitation of *KlPDC1*, a single-gene based system. In: *Communicating Current Research and Educational Topics and Trends in Applied Microbiology*, 2007, pp. 271–282.
[27] Van Ooyen, A. J. J., Dekker, P., Huang, M., Olsthoorn, M. M. A. et al., Heterologous protein production in the yeast *Kluyveromyces lactis*. *FEMS Yeast Res.* 2006, *6*, 381–392.
[28] Alberti, A., Ferrero, I., Lodi, T., *LYS2* gene and its mutation in *Kluyveromyces lactis*. *Yeast* 2003, *20*, 1171–1175.
[29] Yu, J., Jiang, J., Fang, Z., Li, Y. et al., Enhanced expression of heterologous inulinase in *Kluyveromyces lactis* by disruption of *hap1* gene. *Biotechnol. Lett.* 2010, *32*, 507–512.
[30] Feng, Z., Ren, J., Zhang, H., Zhang, L., Disruption of *PMR1* in *Kluyveromyces lactis* improves secretion of calf prochymosin. *J. Sci. Food Agric.* 2011, *91*, 100–103.
[31] Janssen, A., Chen, X. J., Cloning, sequencing and disruption of the *ARG8* gene encoding acetylornithine aminotransferase in the petite-negative yeast *Kluyveromyces lactis*. *Yeast* 1998, *14*, 281–285.
[32] Heinisch, J. J., Buchwald, U., Gottschlich, A., Heppeler, N. et al., A tool kit for molecular genetics of *Kluyveromyces lactis* comprising a congenic strain series and a set of versatile vectors. *FEMS Yeast Res.* 2010, *10*, 333–342.
[33] Rodicio, R., Heinisch, J. J., Yeast on the milky way: Genetics, physiology and biotechnology of *Kluyveromyces lactis*. *Yeast* 2013, *30*, 165–177.
[34] Georis, I., Cassart, J. P., Breunig, K. D., Vandenhaute, J., Glucose repression of the *Kluyveromyces lactis* invertase gene *KlINV1* does not require *Mig1p*. *Mol. Gen. Genet.* 1999, *261*, 862–870.
[35] De Deken, R. H., The Crabtree effect: A regulatory system in yeast. *J. Gen. Microbiol.* 1966, *44*, 149–156.
[36] Förster, J., Famili, I., Fu, P., Palsson, B. Ø. et al., Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* 2003, *13*, 244–253.
[37] Duarte, N. C., Palsson, B. Ø., Fu, P., Integrated analysis of metabolic phenotypes in *Saccharomyces cerevisiae*, *BMC Genomics*, 2004, *5*, 63.
[38] Kuepfer, L., Sauer, U., Blank, L. M., Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Res.* 2005, *15*, 1421–1430.
[39] Nookaew, I., Jewett, M. C., Meechai, A., Thammarongtham, C. et al., The genome-scale metabolic model iIN800 of *Saccharomyces cerevisiae* and its validation: A scaffold to query lipid metabolism. *BMC Syst. Biol.* 2008, *2*, 71.
[40] Herrgård, M. J., Swainston, N., Dobson, P., Dunn, W. B. et al., A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.* 2008, *26*, 1155–1160.
[41] Mo, M. L., Palsson, B. O., Herrgård, M. J., Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.* 2009, *3*, 37.
[42] Österlund, T., Nookaew, I., Bordel, S., Nielsen, J., Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling. *BMC Syst. Biol.* 2013, *7*, 36.
[43] Caspeta, L., Shoaie, S., Agren, R., Nookaew, I. et al., Genome-scale metabolic reconstructions of *Pichia stipitis* and *Pichia pastoris* and in silico evaluation of their potentials. *BMC Syst. Biol.* 2012, *6*, 24.
[44] Sohn, S. B., Graf, A. B., Kim, T. Y., Gasser, B. et al., Genome-scale metabolic model of methylotrophic yeast *Pichia pastoris* and its use for in silico analysis of heterologous protein production. *Biotechnol. J.* 2010, *5*, 705–715.
[45] Chung, B. K., Selvarasu, S., Andrea, C., Ryu, J. et al., Genome-scale metabolic reconstruction and in silico analysis of methylotrophic yeast *Pichia pastoris* for strain improvement. *Microb. Cell Fact.* 2010, *9*, 50.
[46] Sohn, S. B., Kim, T. Y., Lee, J. H., and Lee, S. Y., Genome-scale metabolic model of the fission yeast *Schizosaccharomyces pombe* and the reconciliation of in silico/in vivo mutant growth. *BMC Syst. Biol.* 2012, *6*, 49.
[47] Tarrío, N., Becerra, M., Cerdán, M. E., González Siso, M. I., Reoxidation of cytosolic NADPH in *Kluyveromyces lactis*. *FEMS Yeast Res.* 2006, *6*, 371–380.
[48] Rocha, I., Maia, P., Evangelista, P., Vilaça, P. et al., OptFlux: An open-source software platform for in silico metabolic engineering. *BMC Syst. Biol.* 2010, *4*, 45.
[49] Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G. et al., Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat. Protoc.* 2007, *2*, 727–738.
[50] Kiers, J., Zeeman, A. M., Luttik, M., Thiele, C. et al., Regulation of alcoholic fermentation in batch and chemostat cultures of *Kluyveromyces lactis* CBS 2359. *Yeast* 1998, *14*, 459–469.
[51] Shea, A., Wolcott, M., Daefler, S., Rozak, D. A., Biolog phenotype microarrays. *Methods Mol. Biol.* 2012, *881*, 331–373.

[52] Schomburg, I., Chang, A., Schomburg, D., BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.* 2002, *30*, 47–49.
[53] Dias, O., Gombert, A. K., Ferreira, E. C., Rocha, I., Genome-wide metabolic (re-) annotation of *Kluyveromyces lactis*. *BMC Genomics*, 2012, *13*, 517.
[54] Stelzer, M., Sun, J., Kamphans, T., Fekete, S. P. et al., An extended bioreaction database that significantly improves reconstruction and analysis of genome-scale metabolic networks. *Integr. Biol. (Camb).* 2011, *3*, 1071–1086.
[55] Ma, H., Zeng, A.-P., Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* 2003, 19, 270–277.
[56] Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E. L., Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* 2001, *305*, 567–580.
[57] Smith, T. F., Waterman, M. S., Identification of common molecular subsequences. *J. Mol. Biol.* 1981, *147*, 195–197.
[58] Horton, P., Park, K. J., Obayashi, T., Nakai, K., Protein subcellular localization prediction with WOLF PSORT. *Proceedings of the 4th Asia-Pacific Bioinformatics Conference*, vol. 3, 2006, pp. 39–48.
[59] Greenbaum, D., Colangelo, C., Williams, K., Gerstein, M., Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.* 2003, *4*, 117.
[60] Santos, S. T., *Development of computational methods for the determination of biomass composition and evaluation of its impact in genome-scale models predictions*, Master thesis, Universidade do Minho, 2013.
[61] Von der Haar, T., A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Syst. Biol.* 2008, *2*, 87.
[62] Warner, J. R., The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.* 1999, *24*, 437–440.
[63] Backhaus, K., Heilmann, C. J., Sorgo, A. G., Purschke, G. et al., A systematic study of the cell wall composition of *Kluyveromyces lactis*. *Yeast* 2010, *27*, 647–660.
[64] Tanabe, M., Kanehisa, M., Using the KEGG database resource. *Curr. Protoc. Bioinformatics*, 2012, Chapter 1, Unit1.12.
[65] Wolfe, K. H., Shields, D. C., Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 1997, *387*, 708–713.
[66] Famili, I., Forster, J., Nielsen, J., Palsson, B. O., *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl. Acad. Sci. USA* 2003, *100*, 13134–13139.
[67] Snoek, I. S. I., Steensma, H. Y., Why does *Kluyveromyces lactis* not grow under anaerobic conditions? Comparison of essential anaerobic genes of *Saccharomyces cerevisiae* with the *Kluyveromyces lactis* genome. *FEMS Yeast Res.* 2006, *6*, 393–403.
[68] Mahadevan, R., Schilling, C. H., The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* 2003, *5*, 264–276.
[69] Jouhten, P., Rintala, E., Huuskonen, A., Tamminen, A. et al., Oxygen dependence of metabolic fluxes and energy generation of *Saccharomyces cerevisiae* CEN.PK113-1A. *BMC Syst. Biol.* 2008, *2*, 60.
[70] Boubekeur, S., Camougrand, N., Bunoust, O., Rigoulet, M. et al., Participation of acetaldehyde dehydrogenases in ethanol and pyruvate metabolism of the yeast *Saccharomyces cerevisiae*. *Eur. J. Biochem.* 2001, *268*, 5057–5065.
[71] Dias, O., Rocha, I., Systems Biology in Fungi. In: Paterson, R. (Ed.), *Molecular Biology of Food and Water Borne Mycotoxigenic and Mycotic Fungi*, CRC Press, Boca Raton, (in press).
[72] Verho, R., Richard, P., Jonson, P. H., Sundqvist, L. et al., Identification of the first fungal NADP-GAPDH from *Kluyveromyces lactis*. *Biochemistry* 2002, *41*, 13833–13838.
[73] Jacoby, J., Hollenberg, C. P., Heinisch, J. J., Transaldolase mutants in the yeast *Kluyveromyces lactis* provide evidence that glucose can be metabolized through the pentose phosphate pathway. *Mol. Microbiol.* 1993, *10*, 867–876.
[74] Bianchi, M. M., Tizzani, L., Destruelle, M., Frontali, L. et al., The Ôpetite-negative' yeast *Kluyveromyces lactis* has a single gene expressing pyruvate decarboxylase activity. *Mol. Microbiol.* 1996, *19*, 27–36.
[75] Heinisch, J., Kirchrath, L., Liesen, T., Vogelsang, K. et al., Molecular genetics of phosphofructokinase in the yeast *Kluyveromyces lactis*. *Mol. Microbiol.* 1993, *8*, 559–570.
[76] Zeeman, A.-M., Luttik, M. A. H., Pronk, J. T., Dijken, J. P. et al., Impaired growth on glucose of a pyruvate dehydrogenase-negative mutant of *Kluyveromyces lactis* is due to a limitation in mitochondrial acetyl-coenzyme A uptake. *FEMS Microbiol. Lett.* 1999, *177*, 23–28.
[77] Van Urk, H., Postma, E., Scheffers, W. A., van Dijken, J. P., Glucose transport in crabtree-positive and crabtree-negative yeasts. *J. Gen. Microbiol.* 1989, *135*, 2399–2406.
[78] Baumann, K., Carnicer, M., Dragosits, M., Graf, A. B. et al., A multilevel study of recombinant *Pichia pastoris* in different oxygen conditions. *BMC Syst. Biol.* 2010, *4*, 141.
[79] Møller, K., Olsson, L., Piskur, J., Ability for anaerobic growth is not sufficient for development of the petite phenotype in *Saccharomyces kluyveri*. *J. Bacteriol.* 2001, *183*, 2485–2489.
[80] Shi, N. Q., Jeffries, T. W., Anaerobic growth and improved fermentation of *Pichia stipitis* bearing a *URA1* gene from *Saccharomyces cerevisiae*. *Appl. Microbiol. Biotechnol.* 1998, *50*, 339–345.
[81] Chen, M.-T., Lin, S., Shandil, I., Andrews, D. et al., Generation of diploid *Pichia pastoris* strains by mating and their application for recombinant protein production. *Microb. Cell Fact.* 2012, *11*, 91.
[82] Kurtzman, C. P., Fell, J. W., *The yeasts: A taxonomic study*. Amsterdam, Elsevier, 2000.

Supporting information

Additional file 1 – File with the model in SBML format.
www.*merlin*-sysbio.org/files/iOD907.xml

Additional file 2 – File with additional tables in Excel format.
www.*merlin*-sysbio.org/supplemental_material/additional_file_2.xlsx

Table S1. Biomass components other than the proteins, deoxyribonucleotide and ribonucleotide contents (* mol of biomass component.g biomass^{-1}).

Table S2. Average fatty acid composition (* mol of specific fatty acid.mol average fatty acid^{-1}).

Table S3. Average protein composition (* mol amino acid.g biomass^{-1}. Values used in the iMM904 model are also shown for reference).

Table S4. Deoxynucleoside monophosphates contents in the biomass (mol deoxynucleoside.g biomass^{-1}. Values used in the iMM904 model are also shown for reference).

Table S5. Nucleotide contents in the biomass (mol nucleotide.g biomass^{-1}. Values used in the iMM904 model are also shown for reference).

Table S6. Mannan and 1,3-β-D-glucan contents in the cell (* mol of polysaccharide.g biomass^{-1}; ** g polysaccharide.g biomass^{-1}. Values used in the iMM904 model are also shown for reference).

Table S7. In silico formulation of the Verduyn and other media used for simulating *Kluyveromyces lactis* growth in this work. The upper and lower bounds are presented in (mmol $\cdot$ g^{-1} $\cdot$ h^{-1}).

Table S8. Genes that had their annotation updated.

Table S9. UniProt status of the genes used to develop the iOD907.

Table S10. Genes associated to sterols uptake in *S. cerevisiae* and corresponding *K. lactis* homologues.

Table S11. Reactions not from KEGG or not associated to genes in the model.

Table S12. *K. lactis* growth assessment from in vivo experiments (Biolog Phenotype MicroArrays), from the CBS-KNAW catalogue and from in silico simulations (iOD907) using several carbon sources.

Table S13. Analysis of the model response to different maintenance ATP requirements.

Table S14. Comparison of the behavior of the in silico model to the in vivo knockout experiments.

Table S15. Net conversion of the metabolites available in three environmental conditions utilized in this work.

Table S16. Reactions and respective flux for the three environmental conditions utilized in this work.

Additional file 3 – File with additional data in PDF format.

1.1 Model curation protocol
1.2 Carbon sources assessment
1.3 Additional figures
1.4 Additional References