Title

Systematically gap-filling the genome-scale model of CHO cells

Hamideh Fouladiha [1], Sayed-Amir Marashi [1,*], Shangzhong Li [2, 3], Behrooz Vaziri [4], Nathan E. Lewis [2, 3, 5, *]

[1] Department of Biotechnology, College of Science, University of Tehran, Tehran, Iran

[2] Department of Bioengineering, University of California, San Diego, United States

[3] Novo Nordisk Foundation Center for Biosustainability at the University of California, San Diego, United States

[4] Protein Chemistry and Proteomics Laboratory, Biotechnology Research Center, Pasteur Institute of Iran, Tehran, Iran

[5] Department of Pediatrics, University of California, San Diego, United States

* **Correspondence**:

Sayed-Amir Marashi, Department of Biotechnology, University of Tehran, Enghelab Avenue, Tehran 1417614411, Iran

Nathan E. Lewis , Biomedical Research Facility 2 (BRF2), University of California, CA 92093, San Diego, United States

E-mail addresses:

marashi@ut.ac.ir, nlewisres@ucsd.edu

Abbreviations:

CHO, Chinese hamster ovary; GEM, Genome-scale metabolic model.

Keywords

CHO cells; Constraint-based modeling; Gap-filling; Metabolic network models; Systems biology.

Abstract

Chinese hamster ovary (CHO) cells are the leading cell factories for producing recombinant proteins in the biopharmaceutical industry. In comparison to other mammalian cell types, *in vitro* handling of CHO cells is relatively easy. For example, CHO cells can grow in suspension and reach high cellular densities in bioreactors. Therefore, studying the metabolism of CHO cells to improve the bio-production of these cells is an important subject of research. In this regard, constraint-based metabolic models are useful platforms to perform computational analysis of cell metabolism. Here, we expanded the existing model of Chinese hamster metabolism (*i*CHO1766) with the help of four gap-filling approaches, leading to the addition of 773 new reactions and 335 new genes. We incorporated these into an updated genome-scale metabolic network model of CHO cells, named *i*CHO2101. This updated model substantially increased the number of reactions capable of carrying flux. The addition of these new data provides an important step towards more complete metabolic models of CHO cells.

1. Introduction

Genome-scale metabolic network models (GEMs) are mathematical models that summarize all data about genes, proteins, and reactions known to be involved in the metabolism of a specific cell. Using reliable metabolic models, one can perform virtual experiments in a rapid and inexpensive manner (1, 2). Therefore, GEMs can be helpful tools in cell biology and metabolic engineering by predicting the metabolic state of cells under certain growth conditions (3).

Chinese hamster ovary (CHO) cells are the main workhorse in the biopharmaceutical industry for producing recombinant proteins, such as humanized monoclonal antibodies. These cells were originally obtained from a Chinese hamster (*Cricetulus griseus*) in 1957. Several studies have focused on the optimization of the production of CHO cells using cellular and metabolic engineering methods (4). Experimental manipulation and maintenance of CHO cells, like many other mammalian cell lines, are costly and time-consuming. A reliable metabolic model of CHO cells can be used as a platform to perform computational analyses of cell metabolism to aid in experimental design. Such a model-driven analysis may predict the outcome of experimental tests and reduce the possibility of having false experimental results. Moreover, a CHO metabolic model can be helpful in suggesting genetic engineering and media-design strategies for improving

recombinant protein production (5, 6). Another appreciated application of metabolic models is their role in interpreting "omics" data (7-10). For example, transcriptomic and proteomic data can be mapped to the models to infer new knowledge about the physiological characteristics of cells (11, 12).

One major challenge in the development of genome-scale metabolic network models is our limited knowledge of a cell's metabolism. Specifically, genome-scale metabolic network reconstructions must be iteratively expanded as novel data emerges on enzymes and reactions that occur in the cell of interest. For example, several updates of the GEMs of *Saccharomyces cerevisia* have been published (13), from iND750 (14) and iIN800 (15), to Yeast 5 (16), and ecYeast7 (17). A variety of algorithms have also been developed to predict additional reactions and potential genes that could catalyze such reactions (18), where using machine-learning methods have been helpful (19, 20). These algorithms are particularly useful for expanding the metabolic networks of non-model organisms (21).

Here, we have conducted an in-depth gap-filling of the genome-scale metabolic network reconstruction of the Chinese hamster, *i*CHO1766 (22), and introduce *i*CHO2101, an updated version for enhanced genome-scale modeling of CHO cell metabolism. Compared to the previous version of the CHO model, the number of genes and reactions has been increased, and the numbers of blocked reactions and dead-end metabolites have been reduced by about 10% and 15%, respectively. In other words, more parts of the metabolic model can be active, and more reactions are able to carry fluxes in this new version. These improvements increase the accuracy and precision of the predictions made by the analysis of the metabolic model.

2. Methods

2.1. Analysis of *i*CHO1766

The COBRA toolbox (23) was used for the constraint-based analysis of the metabolic model of CHO cells (*i*CHO1766). Flux Variability Analysis (FVA) (24) was used to find the possible bounds of every flux in steady-state conditions, with no constraints on the flux bounds. If the lower and upper bounds of a specific flux were both equal to zero, that reaction was assumed to be blocked. In the same way, if the upper and lower bound of the exchange flux of a metabolite was zero, that metabolite was considered as a "non-producible and non-consumable" or a "dead-end" metabolite.

2.2. Filling the gaps and validation of the results

In the present study, four independent approaches were used for the gap-filling of *i*CHO1766. The first two approaches were based on automatic gap-filling tools, namely, GapFind/GapFill (25) and GAUGE (26). The GapFind algorithm uses mixed integer linear programming (MILP) to find all metabolites that cannot be produced in steady-state. The "root" gaps are those non-producible metabolites whose filling will unblock the other non-producible (or, "downstream") gaps. Then, the GapFill algorithm selects a minimal subset of reactions from a universal reaction database that must be added to the model in order to convert a non-producible metabolite to a producible one.

In the second approach, we used GAUGE as our computational tool. GAUGE uses transcriptomics data to determine the inconsistencies between genes co-expression and flux coupling in a metabolic model. Then, GAUGE finds a minimal subset of reactions in the KEGG database whose addition can resolve the inconsistencies.

Reactions suggested by GapFind/GapFill and GAUGE (and their associated genes/proteins) were validated before being added to *i*CHO1766 as follows. If the gene ID of the new reaction or the gene ID that is attributed to the enzyme of the new reaction is found in Chinese hamster according to the KEGG database, that new reaction is confirmed. Otherwise, the validation is performed based on the results of BLASTp against the *Cricetulus griseus* (Chinese hamster) transcriptome, using the enzyme of the new reactions and CHO cell transcribed genomic sequences. For each enzyme, in the KEGG database, the amino acid sequences from different species were examined, and the best BLASTp hit was reported. A gene/protein was assumed to be present in Chinese hamster metabolism if a BLAST search hit is found with *e*-value $< 1x10^{-10}$. To have a stricter standard, we only considered hits with query coverage >70%, or, those hits which were of >30% sequence similarity.

Our third gap-filling approach was based on manual assessment of the blocked reactions in *i*CHO1766. In several cases, the absence of an exchange or transport reaction was the cause of reaction blockage in the model. In such cases, we checked if each non-producible or non-consumable metabolite is reported in the Human Metabolome Database (HMDB) (27). If the blocked metabolite was reported to be present in any of the human biofluids (including blood, saliva, and urine), it was assumed that the transport of the metabolite across extracellular membrane of a typical mammalian cell is possible, and therefore, an exchange reaction of that metabolite was added to the model with a high confidence score. If a metabolite was "expected" to be present in biofluids by HMDB, the exchange reaction of that metabolite was added to the model with a low confidence score.

In the fourth approach, the BiGG database (28) was used to retrieve all known biochemical reactions and their corresponding enzymes. Then, the KEGG database was queried to extract the full list of Chinese hamster genes and their association with biochemical enzymes. The intersection of these two lists was considered as the list of potential reactions. Then, the 1766

genes that were present in *i*CHO1766 were subtracted from the list of potential reactions to find those CHO reactions that have counterparts in BiGG, but are not present in *i*CHO1766.

2.3. Analysis of *i*CHO2101

The COBRA toolbox (23) was used for performing flux balance analysis (FBA) and flux variability analysis (FVA) of the updated CHO model when uptake fluxes were unconstrained/constrained. In the unconstrained state, no restrictions were applied to the flux bounds. In the constrained state, on the other hand, only the metabolites of the cell culture medium were allowed to be imported to the model, with a limited flux as defined in *i*CHO1766 (22). Here, FBA was used to predict the maximum growth rate, and FVA was used to calculate possible flux bounds of each reaction while maintaining the maximum growth rate. The reactions with non-zero flux bounds in FVA were considered as "active" reactions.

2.3.1. Gene expression analysis

In order to evaluate the new version of the model and compare it with *i*CHO1766, the transcriptomic data of CHO cells were used. These normalized data include expression levels of more than 23000 genes of CHO-S and CHO-K1 cells across 191 different samples, including published data (22, 29) and unpublished data sets. Data were processed as follows: FastQC v11.1 (30) was used to assess read quality. Trimmomatic v0.33 (31) was used to trim reads with adapters or low-quality scores. STAR2.4.2a (32) was used to align trimmed reads to the CHO-K1 genome (33), followed by calculating fpkm using cufflinks (v2.2.1).

To represent the expression of each gene, the average expression was computed across all 191 samples. The expression of a single-gene reaction was assumed to be proportional to its gene expression. In case of reactions associated with multiple genes, we restricted our analysis to those reactions whose genes were linked either with "OR" or "AND". If all genes of a reaction were linked by "OR"/"AND", the maximum/minimum amounts of gene expressions were attributed to that reaction. Then, we assessed expressions of the reactions in a pathway and compared it with the percentage of blocked reactions in that pathway.

3. Results

3.1. A quarter of reactions in *i*CHO1766 are blocked

The community-consensus genome-scale metabolic models of CHO cells, *i*CHO1766, includes 1766 genes, 6663 reactions, and 4455 metabolites. Using constraint-based modeling (see

Methods), one can observe that about 23% of the reactions (1503 reactions out of the total 6663 reactions) of *i*CHO1766 are blocked. These blocked reactions cannot carry a non-zero flux in steady-state conditions. The reactions of *i*CHO1766 are categorized in 125 metabolic pathways, of which 83 pathways include ten or more reactions. Among these, there are 16 pathways in which at least 50% of the reactions are blocked (Table 1). The distribution of blocked reactions in all metabolic pathways has been shown in Supplementary Table 1. In addition, about 21% of the metabolites (955 metabolites out of total 4455 metabolites) in *i*CHO1766 are "dead-end" metabolites, *i.e.*, they cannot be produced nor consumed in steady-state. These metabolites belong to different subcellular parts of the model (Table 2).

These blocked reactions and dead-end metabolites suggest that *i*CHO1766 includes metabolic gaps (34), which is common in genome-scale metabolic models. Other gaps may also exist in the model, all of which may result in the inconsistencies between model predictions and experimental results. In other words, gaps may decrease the reliability of phenotypic predictions of a metabolic model. Several gap-filling methods have been designed to find the gaps and predict the ways of removing them from the model (35). The majority of these methods use a comprehensive dataset of all known biochemical reactions, which is often obtained from the KEGG database (36). These methods try to find a subset of reactions to be added to the model to fill the gaps and improve model predictions. Gap-filling methods can be classified into three groups. The first group consists of solely-computational methods, which use different computational algorithms and linear or mixed integer linear programming (MILP) to fill the gaps of a model. GapFind/GapFill (25), BNICE (37), FBA-Gap (38), MetaFlux (39), FastGapFill (40), and FastGapFilling (40) are some examples of the first group of methods. The second group of gap-filling methods is phenotype-based methods. These methods take advantage of phenotypic data of the cells, such as viability on different carbon or nitrogen sources, to acquire new data regarding the biochemical reactions of the cell and fill the gaps of the metabolic model of the cell. Smiley (41), GrowMatch (42), OMNI (43), and MinimalExtension (44) belong to the second group. All methods that use various kinds of omics data to fill the gaps of a metabolic model are in the third group, *e.g.*, Sequence-based (45) and Likelihood-based (46) methods, Mirage (47), and GAUGE (26).

In the present study, we decided to use GapFind/GapFill and GAUGE methods to fill the gaps of *i*CHO1766. The results of these two methods were manually validated and added to the model. Besides, two manual gap-filling approaches have been used (see Methods). In the end, representing statistics of the new model and mapping gene expression data will indicate significant improvements in CHO metabolic model.

3.2. Gap filling approaches

Two automatic approaches, namely, GapFind/GapFill and GAUGE, and two manual approaches, were used to fill the gaps of *i*CHO1766. The GapFill method suggested the addition of 121 reactions to the model in order to enable 123 metabolites to be producible (listed in Supplementary Table 2). Some of these 121 reactions can make more than one metabolite to be producible. We validated the predictions by manually searching the KEGG database and also using BLASTp. For example, 4-coumarate (C00811) was a 'root' gap in *i*CHO1766 (a non-producible metabolite in steady-state). In addition, caffeate (C01197) can only be produced from 4-coumarate, and therefore, caffeate was a 'downstream' gap. A reaction (R00737), which is catalyzed by tyrosine ammonia-lyase, can fill both of the aforementioned gaps by transforming tyrosine to 4-coumarate and ammonia. The possibility of tyrosine ammonia-lyase expression in CHO cells was approved using the BLASTp method and therefore, R00737 was added to the model. In total, the addition of 56 reactions was validated, which enabled 87 metabolites to be producible in *i*CHO1766 (Table 3). These new reactions were associated with 30 new genes, which were added to the latest version of the model.

Using the GAUGE method, the inconsistencies between gene co-expression and flux coupling relation of 146 gene pairs were found. GAUGE also suggested solutions for removing the inconsistencies of 64 pairs of them (listed in Supplementary Table 3). Only 37 out of 64 pairs had validated reactions as solutions. In total, 29 reactions were added to *i*CHO1766 using the GAUGE method (Table 4). These new reactions were associated with 3 new genes, which were added to the new version of the model.

In the third gap-filling approach, all non-producible and non-consumable metabolites were searched in the HMDB database, and the equivalent IDs were retrieved. If any of the metabolites were detected in human biofluids, the exchange reaction of that metabolite was added to the model with a high level of confidence. This approach added 257 new reactions to the model (a full list of reactions and HMDB IDs are available in Supplementary Table 4). For example, nonanoate was a dead-end metabolite, which was detected in blood, feces, saliva, and sweat (HMDB0000847). The extracellular export of nonanoate enabled a blocked reaction to carry flux in the linoleate metabolic pathway. There was another group of metabolites that were labeled as "expected to be detected in human biofluids" by HMDB. The exchange reactions of 196 metabolites of this group were added to the model with a low level of confidence (a full list of reactions and HMDB IDs are available in Supplementary Table 5).

With a manual assessment of the blocked reactions in *i*CHO1766, we found that there was a lot of repetition of reactions in different subcellular compartments of the model. In other words, these reactions have the same reactants and products, with precisely the same stoichiometric coefficients, but in different subcellular compartments. In such cases, the absence of appropriate transport reactions caused a lot of blocked reactions. There were 178 blocked

repetitive reactions in the *i*CHO1766, which have no genes, which we therefore suggest for deletion in future curation efforts (all such reactions are listed in Supplementary Table 6). Furthermore, if there was a transport reaction for a metabolite in a subcellular part with no genes in *i*CHO1766, the addition of another transport reaction for that metabolite between other subcellular parts of the new version of the model had a high confidence score. These 139 reactions were added to the new model (Supplementary Table 7).

We found 314 new genes in the fourth approach by searching the BiGG and KEGG databases (see Supplementary Table 8). Twelve of these 314 new genes were also predicted by GapFind/GapFill, and 1 out of 314 new genes was also predicted by GAUGE. The addition of these new genes updated the gene association data of 30 reactions of *i*CHO1766 and also caused 42 new reactions to be added to the new model.

3.3. Analysis of *i*CHO2101

Using the four mentioned gap-filling approaches, a total number of 773 new reactions, 335 new genes, and 76 metabolites were added to *i*CHO1766. The new version of *i*CHO1766, which is named *i*CHO2101, has 2101 genes, 7436 reactions, and 4531 metabolites (see Supplementary Table 10). In *i*CHO2101, 58 pathways contain no blocked reactions, and only 5 pathways have more than 50% blocked reactions (Table 5). In addition, the distribution of dead-end metabolites of *i*CHO2101 in different subcellular compartments has been reduced to less than 10% (Table 6). Figure 1 summarizes the improvements made in the current study for the metabolic model of CHO cells by creating a visual comparison of model statistics, blocked reactions, and dead-end metabolites between *i*CHO1766 and *i*CHO2101.

Using FBA after applying our published uptake and secretion constraints, we found the maximum growth rate in the constrained state was similar for *i*CHO1766 and *i*CHO2101 (0.03 h^{-1}). By performing FVA in the constrained state of *i*CHO1766 and *i*CHO2101, we found the number of “active” reactions in each metabolic pathway had been significantly improved in the gap-filled version of the model. Figure 2 shows the percent of activities of fluxes in 14 metabolic pathways with more than 5 reactions, where the changes between *i*CHO1766 and *i*CHO2101 are more than 30%. For example, all reactions of 'sphingolipid metabolism' are “active” in modeling the growth using *i*CHO2101, thus enabling the analysis of this process, which has been previously reported to be of importance for the growth of CHO cells (48).

3.3.1. Gene expression analysis

We subsequently analyzed the expression of the genes in the metabolic models in 191 RNA-Seq samples. We computed the expression levels of reactions (see Methods). Then, considering the expressions of reactions in the metabolic pathways of the *i*CHO1766, it was revealed that some

of the pathways with a high level of expression had a high percent of blockage. For example, 'androgen and estrogen synthesis and metabolism' had the highest level of expression among blocked pathways, where 98% of the reactions were blocked. In the new model, only 56% of the reactions in the mentioned pathway are still blocked. In another example, 'glyoxylate and dicarboxylate metabolism,' 'methionine and cysteine metabolism,' and 'galactose metabolism' are among the top ten highly expressed pathways, while about 30% of the reactions are blocked in the pathways in *i*CHO1766. In *i*CHO2101, the blocked reactions of the three pathways have been reduced to 11%, 15%, and 7%, respectively. A full list of the pathways and expression levels is available in Supplementary Table 9.

4. Discussion

In the present study, four approaches were used to fill the gaps of *i*CHO1766. At first, we used GapFill that successfully filled 12% (124 out of 1049) of no-production metabolites. Then, using GAUGE, 40% (28 out of 71) of the inconsistencies between genes co-expression and flux coupling relations of reaction pairs were fixed. Furthermore, exchange and transport reactions of the model were revised, using HMDB database. Finally, new genes were added to the model based on KEGG and BiGG databases. All newly predicted reactions and metabolites were subsequently added to the model to generate a new version of the CHO metabolic model, named *i*CHO2101. In total, the percentage of blocked reactions was 21.6% (1441 out of 6663) in *i*CHO1766, which has been reduced to 11.3% (837 out of 7336) in *i*CHO2101. In addition, the percentage of dead-end metabolites from 21.4% (955 out of 4456) in *i*CHO1766 has been reduced to 6.6% (298 out of 4531) in *i*CHO2101. The addition of these new reactions, metabolites, and genes can increase the scope of pathways that can be simulated in CHO cells, and increase the reliability of the model predictions in general for CHO cells with more comprehensive models of CHO cell metabolism.

The importance of CHO cells in the pharmaceutical industry producing recombinant protein drugs is evident. In this regard, a constraint-based metabolic model can be beneficial to have an *in silico* platform to mechanistically model the metabolism of CHO cells.

Bioprocess optimization of CHO cells has been a major topic of research, including studies which focused on the design of compositions of cell culture media (49, 50). Mammalian cell culture media are mostly composed of amino acids. Amino acid metabolism greatly influences the viability and production of CHO cells (51). The average percentage of blocked reactions in the metabolic pathways of different amino acids was reduced from 34.10% in *i*CHO1766 to 13.56% in *i*CHO2101. Therefore, the applicability of CHO model in bioprocess studies can be increased by refining the metabolic models. In addition, with more active metabolic pathways

and more precise gene-protein-reaction associations in a GEM of CHO cells, one is able to infer more accurate cell line-specific models. Such models can address the cell specific metabolic signatures of different cell lines for better predicting biopharmaceutical production capabilities (52).

Acknowledgments

This work was facilitated through generous funding from the Novo Nordisk Foundation through Center for Biosustainability at the Technical University of Denmark (NNF10CC1016517).

Supplementary Information:

Supplementary Table 1: The distribution of blocked reactions of *i*CHO1766 in all metabolic pathways.

Supplementary Table 2: The list of new reactions added by using GapFill method.

Supplementary Table 3: The list of new reactions added by using GAUGE method.

Supplementary Table 4: The list of metabolites that were labeled as "detected in human biofluids" in HMDB and the new reactions associated with them.

Supplementary Table 5: The list of metabolites that were labeled as "expected to be detected in human biofluids" in HMDB and the new reactions associated with them.

Supplementary Table 6: The list of blocked repetitive reactions in *i*CHO1766 that have been suggested for deletion.

Supplementary Table 7: The list of new transport reactions that have a similar reaction in a subcellular part with no genes in *i*CHO1766.

Supplementary Table 8: The list of new reactions added by searching the BiGG database.

Supplementary Table 9: The list of the reactions in metabolic pathways and expression levels associated with them, both in *i*CHO1766 and *i*CHO2101.

Supplementary Table 10: The spreadsheet format of *i*CHO2101.

Conflict of Interest

The authors declare no commercial or financial conflict of interest.

5. References

1. Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current status and applications of genome-scale metabolic models. Genome biology. 2019;20(1):1-18.
2. Fouladiha H, Marashi S-A. Biomedical applications of cell-and tissue-specific metabolic network models. Journal of biomedical informatics. 2017;68:35-49.
3. Zhang C, Hua Q. Applications of genome-scale metabolic models in biotechnology and systems medicine. Frontiers in physiology. 2016;6:413.
4. Wells E, Robinson AS. Cellular engineering for therapeutic protein production: product quality, host modification, and process improvement. Biotechnology journal. 2017;12(1):1600105.
5. Calmels C, McCann A, Malphettes L, Andersen MR. Application of a curated genome-scale metabolic model of CHO DG44 to an industrial fed-batch process. Metabolic engineering. 2019;51:9-19.
6. Traustason B, Cheeks M, Dikicioglu D. Computer-Aided Strategies for Determining the Amino Acid Composition of Medium for Chinese Hamster Ovary Cell-Based Biomanufacturing Platforms. International journal of molecular sciences. 2019;20(21):5464.
7. Hyduke DR, Lewis NE, Palsson BØ. Analysis of omics data with genome-scale models of metabolism. Molecular BioSystems. 2013;9(2):167-74.
8. Kildegaard HF, Baycin-Hizal D, Lewis NE, Betenbaugh MJ. The emerging CHO systems biology era: harnessing the 'omics revolution for biotechnology. Current opinion in biotechnology. 2013;24(6):1102-7.
9. Lakshmanan M, Kok YJ, Lee AP, Kyriakopoulos S, Lim HL, Teo G, et al. Multi-omics profiling of CHO parental hosts reveals cell line-specific variations in bioprocessing traits. Biotechnology and bioengineering. 2019;116(9):2117-29.
10. Richelle A, Chiang AW, Kuo C-C, Lewis NE. Increasing consensus of context-specific metabolic models by integrating data-inferred cell functions. PLoS computational biology. 2019;15(4):e1006867.
11. Schaub J, Clemens C, Kaufmann H, Schulz TW. Advancing biopharmaceutical process development by system-level data analysis and integration of omics data. Genomics and Systems Biology of Mammalian Cell Culture: Springer; 2011. p. 133-63.
12. Richelle A, Joshi C, Lewis NE. Assessing key decisions for transcriptomic data integration in biochemical networks. PLoS computational biology. 2019;15(7):e1007185.
13. Castillo S, Patil KR, Jouhten P. Yeast genome-scale metabolic models for simulating genotype–phenotype relations. Yeasts in Biotechnology and Human Health: Springer; 2019. p. 111-33.
14. Duarte NC, Herrgård MJ, Palsson BØ. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. Genome research. 2004;14(7):1298-309.
15. Nookaew I, Jewett MC, Meechai A, Thammarongtham C, Laoteng K, Cheevadhanarak S, et al. The genome-scale metabolic model *iIN800* of *Saccharomyces cerevisiae* and its validation: a scaffold to query lipid metabolism. BMC systems biology. 2008;2(1):71.
16. Heavner BD, Smallbone K, Barker B, Mendes P, Walker LP. Yeast 5–an expanded reconstruction of the *Saccharomyces cerevisiae* metabolic network. BMC systems biology. 2012;6(1):55.
17. Sánchez BJ, Zhang C, Nilsson A, Lahtvee PJ, Kerkhoven EJ, Nielsen J. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Molecular systems biology. 2017;13(8).
18. Karlsen E, Schulz C, Almaas E. Automated generation of genome-scale metabolic draft reconstructions based on KEGG. BMC bioinformatics. 2018;19(1):467.
19. Medlock GL, Papin JA. Guiding the refinement of biochemical knowledgebases with ensembles of metabolic networks and machine learning. Cell Systems. 2020.

20. Medlock GL, Papin J. Medusa: software to build and analyze ensembles of genome-scale metabolic network reconstructions. BioRxiv. 2019:547174.
21. Biggs MB, Papin JA. Managing uncertainty in metabolic network structure and improving predictions using EnsembleFBA. PLoS computational biology. 2017;13(3):e1005413.
22. Hefzi H, Ang KS, Hanscho M, Bordbar A, Ruckerbauer D, Lakshmanan M, et al. A consensus genome-scale reconstruction of Chinese hamster ovary cell metabolism. Cell Systems. 2016;3(5):434-43.
23. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nature Protocols. 2007;2(3):727-38.
24. Burgard AP, Vaidyaraman S, Maranas CD. Minimal reaction sets for *Escherichia coli* metabolism under different growth requirements and uptake environments. Biotechnology progress. 2001;17(5):791-7.
25. Kumar VS, Dasika MS, Maranas CD. Optimization based automated curation of metabolic reconstructions. BMC bioinformatics. 2007;8(1):212.
26. Hosseini Z, Marashi S-A. Discovering missing reactions of metabolic networks by using gene co-expression data. Scientific reports. 2017;7:41774.
27. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic acids research. 2017;46(D1):D608-D17.
28. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic acids research. 2015;44(D1):D515-D22.
29. Van Wijk XM, Döhrmann S, Hallström BM, Li S, Voldborg BG, Meng BX, et al. Whole-genome sequencing of invasion-resistant cells identifies laminin α2 as a host factor for bacterial invasion. MBio. 2017;8(1):e02128-16.
30. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.
31. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
32. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
33. Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X, et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. Nature biotechnology. 2011;29(8):735.
34. Orth JD, Palsson BØ. Systematizing the generation of missing metabolic knowledge. Biotechnology and bioengineering. 2010;107(3):403-12.
35. Pan S, Reed JL. Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. Current opinion in biotechnology. 2018;51:103-8.
36. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic acids research. 2016;45(D1):D353-D61.
37. Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD, Broadbelt LJ. Exploring the diversity of complex metabolic networks. Bioinformatics. 2005;21(8):1603-9.
38. Brooks JP, Burns WP, Fong SS, Gowen CM, Roberts SB. Gap detection for genome-scale constraint-based models. Advances in bioinformatics. 2012;2012.
39. Latendresse M, Krummenacker M, Trupp M, Karp PD. Construction and completion of flux balance models from pathway databases. Bioinformatics. 2012;28(3):388-96.
40. Thiele I, Vlassis N, Fleming RM. fastGapFill: efficient gap filling in metabolic networks. Bioinformatics. 2014;30(17):2529-31.
41. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, et al. Systems approach to refining genome annotation. Proceedings of the National Academy of Sciences. 2006;103(46):17480-4.
42. Kumar VS, Maranas CD. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. PLoS computational biology. 2009;5(3):e1000308.
43. Herrgård MJ, Fong SS, Palsson BØ. Identification of genome-scale metabolic network models using experimentally measured flux profiles. PLoS computational biology. 2006;2(7):e72.
44. Christian N, May P, Kempa S, Handorf T, Ebenhöh O. An integrative approach towards completing genome-scale metabolic networks. Molecular BioSystems. 2009;5(12):1889-903.
45. Krumholz EW, Libourel IG. Sequence-based network completion reveals the integrality of missing reactions in metabolic networks. Journal of Biological Chemistry. 2015;290(31):19197-207.
46. Benedict MN, Mundy MB, Henry CS, Chia N, Price ND. Likelihood-based gene annotations for gap filling and quality assessment in genome-scale metabolic models. PLoS computational biology. 2014;10(10):e1003882.

47. Vitkin E, Shlomi T. MIRAGE: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks. Genome biology. 2012;13(11):R111.
48. Hanada K, Nishijima M, Kiso M, Hasegawa A, Fujita S, Ogawa T, et al. Sphingolipids are essential for the growth of Chinese hamster ovary cells. Restoration of the growth of a mutant defective in sphingoid base biosynthesis by exogenous sphingolipids. Journal of Biological Chemistry. 1992;267(33):23527-33.
49. Ritacco FV, Wu Y, Khetan A. Cell culture media for recombinant protein expression in Chinese hamster ovary (CHO) cells: History, key components, and optimization strategies. Biotechnology progress. 2018;34(6):1407-26.
50. Galbraith SC, Bhatia H, Liu H, Yoon S. Media formulation optimization: current and future opportunities. Current opinion in chemical engineering. 2018;22:42-7.
51. Salazar A, Keusgen M, von Hagen J. Amino acids in the cultivation of mammalian cells. Amino acids. 2016;48(5):1161-71.
52. Carinhas N, Duarte TM, Barreiro LC, Carrondo MJ, Alves PM, Teixeira AP. Metabolic signatures of GS-CHO cell clones associated with butyrate treatment and culture phase transition. Biotechnology and bioengineering. 2013;110(12):3244-57.

Tables:

Table 1. A list of metabolic pathways of *i*CHO1766 that more than 50 percent of the metabolic reactions in that pathway is blocked.

Biochemical Pathway	**Number of Blocked Reactions**	**Total Number of Reactions**	**Percent Blocked Reactions (%)**
Chondroitin synthesis	45	45	100
Linoleate metabolism	14	14	100
Selenoamino acid metabolism	16	16	100
Vitamin E metabolism	23	23	100
Xenobiotics metabolism	25	25	100
Arachidonic acid metabolism	72	73	98.63
Androgen and estrogen synthesis and metabolism	49	50	98
Eicosanoid metabolism	212	244	86.89
N-glycan biosynthesis	64	77	83.12
Vitamin D metabolism	22	29	75.86
Miscellaneous	45	69	65.22
Vitamin C metabolism	8	14	57.14
Tyrosine metabolism	59	106	55.66
Tryptophan metabolism	36	66	54.55
Glycosphingolipid metabolism	7	13	53.85
Urea cycle	33	63	52.38

Table 2. The distribution of dead-end metabolites of *i*CHO1766 in each subcellular part.

Subcellular part	**Total number of metabolites**	**Number of the blocked metabolites**	**Percent of the dead-end metabolites (%)**
Extracellular [e]	606	56	9.24
Cytoplasm [c]	1652	323	19.55

Endoplasmic reticulum [r]	479	153	31.94
Mitochondrion [m]	620	171	27.58
Peroxisome [x]	318	133	41.82
Nucleus [n]	158	66	41.77
Lysosome [l]	260	33	12.69
Golgi apparatus [g]	361	20	5.54

Table 3. New validated reactions predicted by using the GapFill method to be added to the model. The numbers in parenthesis are query coverage, *e*-value, and sequence similarity, respectively.

Blocked Metabolite ID	Predicted Reaction's KEGG ID	KEGG Gene ID (for the Predicted Reaction)	Blast result (If needed, in case of no gene or enzyme KEGG ID)	Comments
3deccrn[c], C05264[c], C05264[m]	R03778	100753943, 100754813, 100758239, 100765829	✓	
	R04743	100754698, 100757947, 100761491	✓	
ak2gchol_cho[c], ak2gp_cho[c], and ak2gpe_cho[c], dak2gpe_cho[c], C03201[c], C03715[c]	R04311	100756809	✓	
	R05190	-	ERE79474.1 (acyl-CoA synthetase family member 3) with WP_012013866 = (95% 3e-36 28%)	
	R10104	100758702	✓	
C00243[l]	R01100	100766856, 100767446	✓	3.2.1.108 : 100766856; 3.2.1.23 : 100767446
C00247[c], C01507[c]	R02925	-	EGW06281.1 (Carbonyl reductase [NADPH] 2) with WP_011337990 = (98% 5e-25 29%)	
C00257[c]	R01738	-	EGW06281.1 (Carbonyl reductase [NADPH] 2) with YP_002410598 = (98% 5e-34 32%)	1.1.1.69
	R01740		EGW06281.1 (Carbonyl reductase [NADPH] 2) with WP_011565275 = (97% 7e-34 32%)	1.1.69
C00265[c]	R03511	100755703, 100757934, 103162274	✓	
	R05830		ERE75446.1 (vitamin K epoxide reductase complex subunit 1-like protein 1) with Q8N0U8 = (72% 5e-71 84%)	
C00309[c]	R01895	-	EGW01280.1 (Dehydrogenase/reductase SDR family member 7B) with WP_015365771 = (93% 2e-26 33%)	
C00437[c]	R10466		ERE85082.1 (arginase-1-like protein) with D2Z025 = (75% 5e-09 27%)	
C00461[c]	R01206	100750633, 100750757, 100760661, 103161867, 103161868, 103163420	✓	
C00499[c], C01551[c], C11821[c], C12248[c]	R02106	100768251	✓	
C00811[c], C01197[c]	R00737	-	ERE91835.1 (histidine ammonia-lyase-like protein) with NP_719898 = (91% 8e-90 38%)	

C00988[c]	R04620	100750903, 100751196,100751774, 100771587, 103159036, 103159088	✓	
C01083[c]	R01557	-	ERE74261.1 (neutral and basic amino acid transport protein rBAT-like protein) with WP_002548616 (83% 4e-66 32%)	
C01127[c]	R04445	100764994	✓	
C01176[c], C05138[c]	R08516	100758683	✓	4.1.2.30 : 100758683
C01189[c]	R07215	100752960	✓	
C01241[c], C04308[c]	R02056	100767954	✓	
C01528[c], C05172[c]	R03595	100770125, 100775017 AND	✓	
	R04620	100750903 , 100751196 , 100751774 , 100771587 , 103159036 , 103159088	✓	
C01601[c], C04717[c], C08261[c], CE2006[c], CE2576[c], CE2577[c], CE6504[c], CE6506[c]	R03626	-	ERE67202.1 (arachidonate 5-lipoxygenase) with XP_002516771 = (70% 7e-46 27%	
C01802[c], C05107[c]	R07507	100769920	✓	
C02576[c], peracd[c]	R03945	-	ERE88510.1 (alcohol dehydrogenase 6-like protein) with A0A0K2YIV5 = (97% 9e-43 31%) AND EGW05976.1 (Quinone oxidoreductase) with A0A084FZJ5 = (93% 2e-38 34%)	
C03366[c]	R04620	100750903, 100751196,100751774, 100771587, 103159036, 103159088	✓	
C03681[c], C13712[c], CE5072[c]	R02208	100761447, 100766917, 100770660	✓	
C03845[c]	R07215	100752960	✓	
C04722[r]	R04807	100751584	✓	
C04805[c]	R07034	100751356, 100756109, 100764638, 100766519, 100766810, 100771188, 100775000	✓	
C04853[c], CE2056[c], CE3554[c]	R03866	100755384, 100761725, 100773031, 100773326, 100774306, 100774594	✓	
	R08516	100758683	✓	4.1.2.30 : 100758683
C05141[c], C05504[c]	R03089	100751269	✓	
C05141[r], C05504[r]	R04681	100750866, 100751291, 100751897, 100762147, 100766230, 100767580	✓	
C05638[c] C05639[c] C05651[c]	R04911	100773211	✓	
C05688[c]	R03599		ERE85901.1 (selenocysteine lyase) with NP_057926 = (99% 0.0 90%)	
	R07933	100757464	✓	
C05691[c]	R04620	100750903, 100751196, 100751774, 100771587, 103159036, 103159088	✓	
C05698[c], C05699[c]	R04941	-	EGW06584.1 (Cystathionine gamma-lyase) with WP_002493862 = (92% 2e-85 40%)	
C05768[c] C05769[c]	R03166		✓	R03166 is a spontaneou

				s reaction
	R04972	100753284	✓	
C05839[c], C06738[c]	R04998	100757820	✓	
	R04999		✓	R04999 is a non-enzymatic reaction
C05947[c]	R04444	100764994	✓	
C06128[c]	R04018	100689090, 100689301, 100689373, 100774175	✓	
C06133[c]	R03354	100689438	✓	
	R03488	100754838	✓	
	R04583	100768920	✓	
C06178[c]	R01153		EGW12892.1 (Spermidine synthase) with XP_002534321 = (64% 5e-76 50%)	
	R01920	100756588	✓	
	R04027	100762635, 100762926	✓	
C06196[c]	R04620	100750903, 100751196, 100751774, 100771587, 103159036, 103159088	✓	
	R10235	100768978, 100769259	✓	
C14825[c], CE2047[c]	R07055	100751762, 100752064, 100753681, 100754177, 100754462, 100755851, 100756757, 100764171, 100764471, 100764768, 100765057, 100765891, 100766524, 100767391, 100772776, 100773059, 100773351	✓	
C14826[c], CE2049[c]	R07056	100751762, 100752064, 100753681, 100754177, 100754462, 100755851, 100756757, 100764171, 100764471, 100764768, 100765057, 100765891, 100766524, 100767391, 100772776, 100773059, 100773351	✓	
C15610[c]	R08726	100751584	✓	
	R08727	100754734	✓	
C16216[c], C16217[c]	R07758	100763438	✓	
CE1292[c], CE1298[c]	R01463 AND R08505	100689275 AND 100751584	✓	
CE2084[c] CE5815[c] CE7096[c]	R07034	100751356, 100756109, 100764638, 100766519, 100766810, 100771188, 100775000	✓	

Table 4. New validated reactions predicted by using the GAUGE method to be added to the model.

Predicted Reaction ID	**KEGG Gene ID (for the Predicted Reaction)**	**Blast results (If needed, in case of no gene or enzyme KEGG ID)**	**Comments**
R00270	-	✓	non-enzymatic hydrolysis
R00524	-	ERE88882.1 (bis(5'-adenosyl)-triphosphatase-like protein) with XP_002410848 = (20% 6e-23 52%)	3.5.1.49
R00310	100764152	✓	
R00557, R00558	100758127, 100760062, 100769961	✓	
R00648	100753207, 100755812, 100773063	✓	

R01658, R02003	100754883, 100767405	✓	
R02061	100767405	✓	
R02285	-	EGV96886.1: (Agmatinase, mitochondrial) with WP_057563128 = (95% 4e-10 23%,)	3.5.3.8
R03189	100754097	✓	
R03222	100767777	✓	
R03326	100767691	✓	
R04283	-	✓	multi-step reaction, non-enzymatic, incomplete reaction
R04666	100762944	✓	
R06127, R06128	100765573, 100771009	✓	
R06238	100768412	✓	
R06895	-	XP_003501431.1 (radical S-adenosyl methionine domain-containing protein 1, mitochondrial isoform X1) with WP_057908418 = (85% 4e-58 35%)	1.3.99.22
R07267, R09250, R09251	-	EGV97845.1 (Decaprenyl-diphosphate synthase subunit 1) with V5V4V5 = (97% 1e-71 41%)	2.5.1.84
		EGV97845.1 (Decaprenyl-diphosphate synthase subunit 1) with XP_010697478 = (97% 4e-59 35%)	2.5.1.85
R07364	100754671, 100765075	✓	
R07396	100754678	✓	
R08892, R10130	-	ERE70900.1 (sorbitol dehydrogenase) with Q2MF72 = (94% 2e-22 26%)	1.1.1.329
R09248	-	EGV97845.1 (Decaprenyl-diphosphate synthase subunit 1) with = WP_001513338 (97% 4e-41 30%)	2.5.1.90
R10107	-	EGV91790.1 (Nitric oxide synthase, endothelial) with O34453 = (98% 2e-98 43%)	1.14.13.165
R10221	100765199	✓	

Table 5. A list of metabolic pathways of *i*CHO2101 that more than 50 percent of the metabolic reactions in that pathway is blocked.

Biochemical Pathway	**Number of Blocked Reactions**	**Total Number of Reactions**	**Percent Blocked Reactions (%)**
Xenobiotics metabolism	25	25	100
Selenoamino acid metabolism	15	21	71.43
Androgen and estrogen synthesis and metabolism	29	51	56.86
Arachidonic acid metabolism	42	74	56.76
Eicosanoid metabolism	127	245	51.84

Table 6. The distribution of dead-end metabolites of *i*CHO2101 in each subcellular part.

Subcellular part	Total number of metabolites	Number of the dead-end metabolites	Percent of the dead-end metabolites (%)
Extracellular [e]	609	2	0.33
Cytoplasm [c]	1717	156	9.08
Endoplasmic reticulum [r]	482	45	9.34
Mitochondrion [m]	625	52	8.32
Peroxisome [x]	318	26	8.18
Nucleus [n]	159	12	7.55
Lysosome [l]	260	3	1.15
Golgi apparatus [g]	361	2	0.55

Figures:

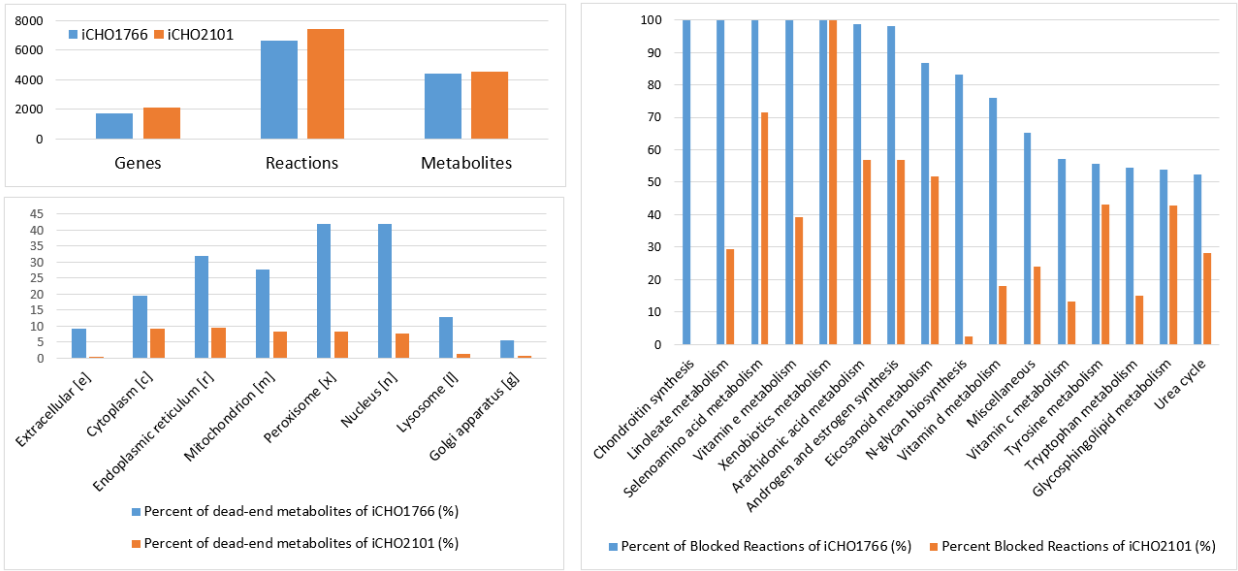

Figure 1.

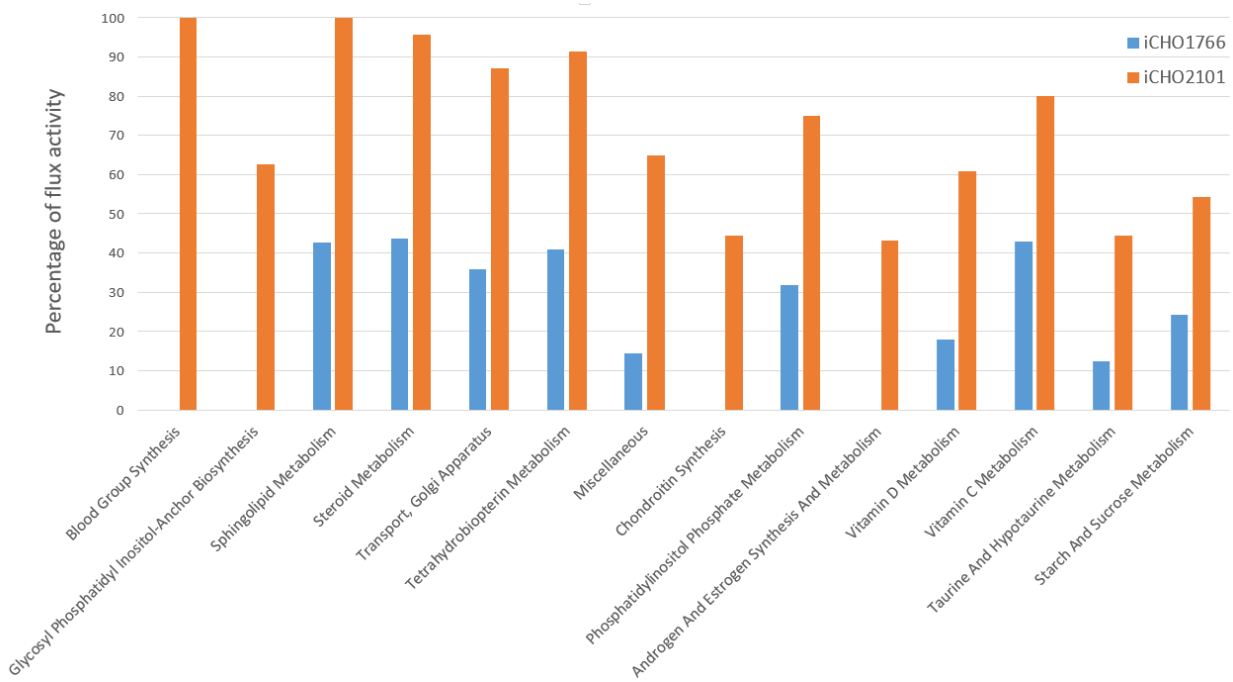

Figure 2.

Figure Legends:

Figure 1. A visual comparison between the statistics of *i*CHO1766 and *i*CHO2101. Part (A) shows the number of genes, reactions, and metabolites. Part (B) shows the distribution of dead-end metabolites in different subcellular parts. Part (C) shows the percent of blockage in the selected pathways reported in Table 1.

Figure 2. A visual comparison of flux activities in 14 metabolic pathways of *i*CHO1766 and *i*CHO2101, where the changes are more than 30% in comparison.