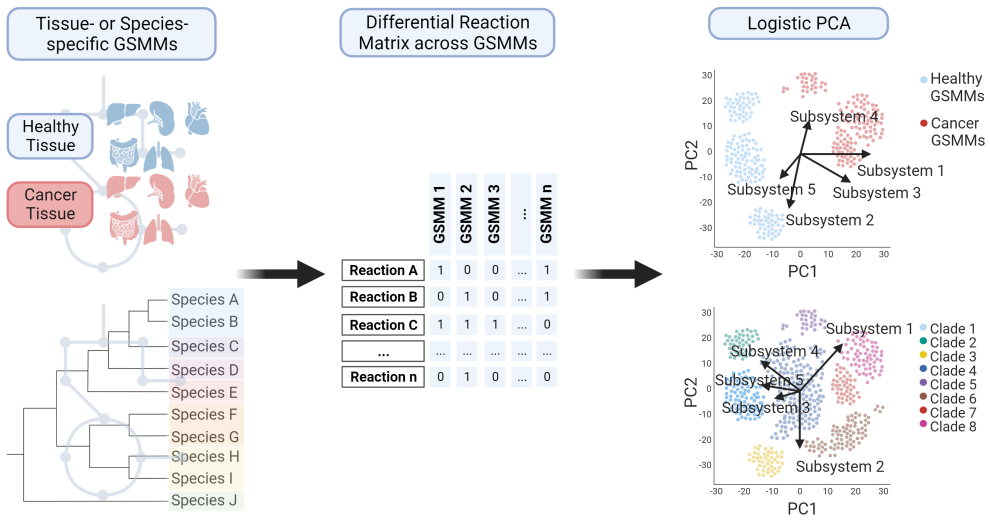
Differential logistic PCA for large-scale comparison of GSMMs and pathway analysis

Leopold Zehetner[1,2,3], Diana Széliová[1], Barbara Kraus[3], Juan A. Hernandez Bort[3], and Jürgen Zanghellini[1*]

1 Department of Analytical Chemistry, Faculty of Chemistry, University of Vienna, Vienna, Austria
2 Vienna Doctoral School in Chemistry (DoSChem), University of Vienna, Vienna, Austria
3 Gene Therapy Process Development, Baxalta Innovations GmbH, a Part of Takeda Companies, Orth an der Donau, Austria

* juergen.zanghellini@univie.ac.at

Graphical Abstract



Abstract

Genome-scale metabolic models (GSMMs) provide a platform for metabolic analyses at the systems level by mathematically depicting biochemical reactions. Recent advances led to a large amount of GSMMs for different species or tissue-specific reconstructions from global GSMMs. Currently, the comparison of GSMMs mostly depends on methods like dimension reduction algorithms, distance matrices, or predefined environmental conditions. However, these approaches are either hard to interpret or rely on subjective assumptions. Here we introduce a novel use of logisitic principal component analysis (lPCA) to simplify the comparison of GSMMs. LPCA is especially suited for binary data, derived from the presence or absence of reactions across GSMMs. Its advantage is

its ability to cluster GSMMs while pinpointing the reactions that influence the categorization. When we applied lPCA to three distinct datasets, it effectively highlighted phylogenetic relationships and distinctly separated the GSMMs of healthy tissues from those of cancer tissues. Furthermore, the identified subsystems and their associated reactions that influenced the clustering align with existing knowledge.

Introduction

GSMMs are mathematical representations of species- or tissue-specific metabolic reactions [1] and have been successfully applied to study diseases [2–4], optimize bioprocesses [5–7], or investigate metabolic differences across species [8–10]. The increasing interest in GSMMs and the invention of tools to automatically generate GSMMs from genomes *de novo* [11–13], or reconstruct tissue-specific GSMMs by integrating omics data [14–16], led to a large amount of GSMMs, each consisting of several hundreds to thousands of reactions. Large numbers of GSMMs can be efficiently compared by clustering following dimension reduction algorithms or differential pathway analysis. The sheer size of data requires the development of automated tools for comparison and pathway analysis.

Current methods to determine heterogeneity in GSMMs include various distance metrics, such as Jaccard coefficient [17], and dimensionality reduction algorithms, such as principal component analysis (PCA) [8] or t-distributed stochastic neighborhood embedding (t-SNE) [9, 18]. These methods have been regularly applied to GSMM-derived data, including binary matrices [8, 9], or simulated phenotypes among different environmental states [8].

Binary matrices summarise GSMMs by assigning 1 if a reaction is present in a GSMM, or 0 if the reaction is absent. In previous studies, these binary matrices have been analyzed using Jaccard coefficient [17] or t-SNE clustering [9, 18]. Jaccard coefficient can be used as a measure of similarity between two GSMMs. Visualizing pairwise Jaccard coefficients in heatmaps allows the identification of clusters of similar GSMMs [17]. However, this approach does not provide insights into specific pathways responsible for the heterogeneity between GSMMs. Binary data have been efficaciously subjected to t-SNE analysis [9, 18], yielding consistent clustering of GSMMs. However, t-SNE mandates specific prerequisites, such as distance metrics, and hyperparameters, which might result in erroneous clustering outcomes [19–21]. Additionally, the inherent non-deterministic nature of t-SNE poses reproducibility concerns. Another limitation of t-SNE is that there is no straightforward way to identify which variables drive the clustering. In contrast, PCA allows for calculation of loadings - reactions or genes that contribute most to the principal components. These can then be used for pathway analysis.

Alternatively, dimensionality reduction techniques can be applied to the results of simulations, such as flux balance analysis [22]. The flux distributions or growth rates can be predicted for different environmental conditions and serve as input for PCA. PCA of growth rates was used to suggest potential auxotrophies [8, 10]. While this approach has become a well established method for comparing GSMMs, it necessitates the incorporation of preassumed environmental data. Another challenge is that the predicted growth rates can be similar even in different environments, which might lead to a reduced discriminative capacity. One potential remediation is to focus solely on simulated growth rates from selected environmental conditions exhibiting significant variation across GSMMs; however, this mandates the integration of subjective parameters or thresholds.

To overcome the limitations of existing methods, we use lPCA for GSMMs classification. LPCA is an adaption of classic PCA to analyze heterogeneity in binary

data. In a classic PCA, continuous data is transformed to a new coordinate system such that the greatest variance lies on the first coordinate. In contrast, lPCA extends this approach to binary data by incorporating a logistic link function, allowing for the modeling of non-linear relationships between the original variables. While PCA is unsuitable for analyzing binary data due to its reliance on variance measures, which could be less meaningful for binary values, lPCA is specifically tailored to analyze binary data [23]. The detailed computation is beyond the scope of this application note, but the interested reader might be referred to implementations [24, 25], especially to the one from Landgraf and Lee [23]. LPCA has been used with other biological data, such as binary genomics data [25], but it has not been explored for binary data derived from GSMMs yet. LPCA applied to GSMMs allows efficient clustering only based on the presence or absence of reactions in GSMMs as well as the identification of reactions that contribute the most to clustering, which facilitates pathway analysis. In this study, we used lPCA to the classification of GSMMs, highlighting its benefits compared to existing methods. We validate our approach by recovering the clusters identified by PCA of transcriptomes and reconstruct phylogenetic associations. Furthermore, we identify the key pathways that differentiate the clusters of GSMMs. Overall, we provide an alternative for streamlined pathway analysis that elucidates variations among GSMMs.

Methods

Data collection and pre-processing

Genome-scale metabolic model

We acquired GSMMs and reconstructions from three distinct sources as outlined below. Utilizing COBRApy (v0.26.3) [26], we generated binary matrices for each data set, indicating the presence (1) or absence (0) of individual reactions within species-specific GSMMs or tissue-specific reconstructions. To refine the data sets, columns consisting solely of ones, indicating conserved reactions across species or tissues, were systematically excluded from the binary matrices. We refer to the resulting data as the differential binary reaction matrix of the pan-GSMM or the pan-reconstruction.

Escherichia. Monk [8] recently reconstructed GSMMs for 222 *Escherichia* species, with the objective of investigating metabolic variations across 570 distinct environmental conditions. The resulting pan-GSMM contains 3342 reactions, all found within at least one strain. 1688 reactions are not uniformly present across all strains.

Fungi. Lu, *et al.* [9] reconstructed GSMMs using sequencing data from 332 fungal genomes [27], augmented by eleven additional fungal GSMMs categorized as "Outgroup". The authors first built a pan genome-scale reconstruction containing all annotated genes across all species. The set of pan-reactions contained 4599 unique reactions and formed the pool of reactions for the individual species-specific GSMMs and gap-filling. 2519 reactions were not found in all species.

For comparison, we repeated a phylogenetic anlysis [27] and a t-SNE analysis [9] as outlined in the original publications.

Human. We utilized normalized gene expression levels derived from 50 healthy tissue samples [28] and 30 cancer tissue samples [29] taken from the Human Protein Atlas [28, 29] to reconstruct tissue-specific genome-scale metabolic reconstruction. Reactions were taken from the latest iteration of the human metabolic model, "Human1" [30]. Gene expression was considered if the normalized transcripts per million

(nTPM) value exceeded 0.2. To avoid method-specific bias from reconstruction or gap-filling, only reactions with gene assignments were used for the tissue-specific reconstruction. The set of pan-reactions contained 7975 unique reactions, of which 2390 reactions were not found in all tissues.

PCA and logisitic principal component analysis (lPCA)

PCA was executed on normalized transcriptomic data sets sourced from both healthy and cancer tissues. Prior to processing, the data was subjected to tissue-specific z-scaling and subsequently analyzed using the R function `prcomp`.

LPCA was performed on the differential binary reaction matrix using the `logisticPCA` function from the "logisticPCA" package (v0.2) [23] in R (v4.3.1) with the natural parameter "m" set to 0. Due to the considerable dimensions of the binary matrices, we opted for partial decomposition. For each specific differential reaction $r_k^{(j)}$, we derived distinct loadings. These loadings were then categorized based on the inherent subsystems j within the collection of GSMMs. That is, every principal component p_i was represented as

$$p_i = \sum_{j=1}^{\#S} \sum_{k=1}^{\#R_j} \alpha_{i,k}^{(j)} r_k^{(j)}, \tag{1}$$

where the first sum traverses all metabolic subsystems indexed by $j \in 1, .., \#S$, and the second sum iterates through all differential reactions $j \in 1, .., \#R_j$ within a given subsystem. Here, $\#S$ represents the total number of metabolic subsystems, and $\#R_j$ indicates the count of differential reactions within a specific subsystem j.

The (average) component of the subsystem-specific loading vectors visualized in Fig 1 to Fig 4 were calculated as

$$\text{avg}\left(\alpha_i^j\right) = \frac{1}{\#R_j} \sum_{k=1}^{\#R_j} \alpha_{i,k}^{(j)}. \tag{2}$$

We used the magnitude of (2) over the first two principal components, i.e.

$$\left|\text{avg}\left(\alpha^j\right)\right| = \sqrt{\sum_{i=1}^{2} \text{avg}\left(\alpha_i^j\right)} \tag{3}$$

for ranking. We also measured the mean length of every subsystem-specific loading vector, which was defined as

$$\text{avg}\left|\alpha^j\right| = \frac{1}{\#R_j} \sum_{k=1}^{\#R_j} \sqrt{\sum_{i=1}^{2} \left(\alpha_{i,k}^j\right)^2}, \tag{4}$$

where the inner sum traverses over the first two principal components.

Both measures were employed to identify key subsystems responsible for driving the observed differentiation.

Phylogenetic analysis

Escherichia. Genomes from 222 *Escherichia* strains were downloaded from Enterobase (v1.1.5) [31]. Next, proteins from genomic sequences were predicted using prodigal (v2.6.3) [32], followed by phylogenetic comparison using OrthoFinder (v2.5.5) [33]. The phylogenetic tree was then plotted using the "ape" package (v5.7-1) [34] in R.

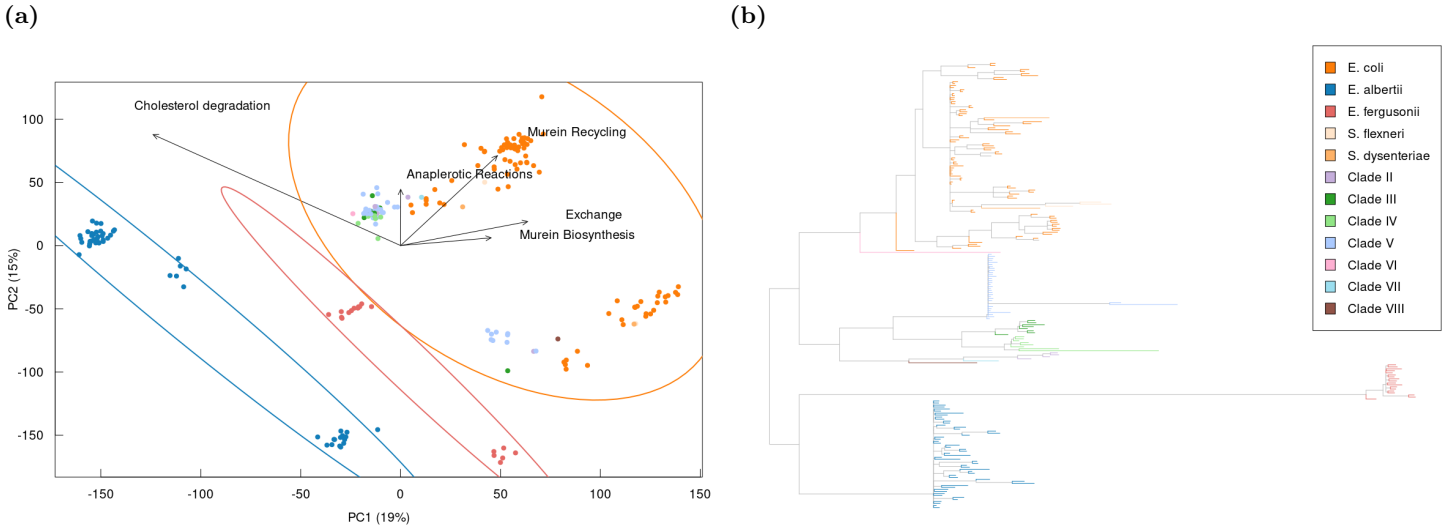


Fig 1. Relationship based on differential reactions across from *Escherichia* GSMMs using lPCA compared to genome-based phylogeny Phylogenetic relations could be reconstructed after lPCA (a) compared to genome-based relationships (b) and showed improved clustering compared to PCA on simulated growth rates [8] (see supplementary Fig S2 a and Fig S2 b). (a) Subsystem-centric loadings were calculated as the mean of all loadings from reactions assigned to the specific pathway. The five most influential subsystems were obtained by calculating the magnitude of the subsystem-centric loadings. (b) Whole genome-based phylogenetic tree of 222 *Escherichia* strains indicating the three major clusters: *E. albertii*, *E. fergusonii* and a supercluster, containing the remaining clades.

Fungi. The phylogenetic tree was reproduced using the "ape" package [34] based on a Newick file published by Shen *et al* [27].

principal component analysis of *Escherichia*

For visual comparison, the PCA plot based on simulated growth rates from environmental conditions was reproduced for *Escherichia* strains from [8]. To obtain a similar clustering, a correlation matrix was computed, and PCA was performed using the `princomp` function in R.

t-SNE analysis of fungi

Reactions from all yeast-specific GSMMs were condensed in a binary matrix using COBRApy [26], followed by Hamming-distance based t-SNE calculation using the R package "Rtsne" (v0.16) [35].

Results and Discussion

Comparison of 222 *Escherichia*-specific GSMMs

We used lPCA to analyze 222 strain-specific GSMMs of the genus *Escherichia* reconstructed by Monk [8], see Fig 1a. LPCA identified seven clusters. Three of them solely associate with *E. albertii* strains (Fig 1a, blue), two of them associated with *E. fergusonii* strains (Fig 1a, red), and the remaining two consist of *E. coli*, *S. dysenteriae*, *S. flexneri*, as well as the Clades II to VIII (Fig 1a, orange). While *E. coli* strains did

Table 1. Top five subsystems driving cluster separation in 222 *Escherichia* strains in lPCA. Subsystems were ranked either according to the magnitude of the average loadings per subsystem for the first two principal components, $\left|\text{avg}\left(\alpha_i^j\right)\right|$, or according to the average magnitude of the loadings per subsystem, $\text{avg}\left|\alpha^j\right|$, see equations (3) and (4), respectively.

	before PRCOA1 reassignment			after PRCOA1 reassignment		
Subsystem j	$\#R_j$	$\lvert\text{avg}\left(\alpha^j\right)\rvert$	$\text{avg}\lvert\alpha^j\rvert$	$\#R_j$	$\lvert\text{avg}\left(\alpha^j\right)\rvert$	$\text{avg}\lvert\alpha^j\rvert$
Murein Recycling	4	0.0430	0.0612	4	0.0430	0.0612
Exchange	1	0.0332	0.0332	1	0.0332	0.0332
Anaplerotic Reactions	4	0.0222	0.0385	4	0.0222	0.0385
Histidine Metabolism	10	0.0211	0.0336	9	0.0148†	0.0287†
Cholesterol degradation	-‡	-‡	-‡	29	0.0759	0.0760
Murein Biosynthesis	3	0.0229	0.0265†	3	0.0229	0.0265†
Nucleotide Salvage Pathway	48	0.0139†	0.0310	48	0.0139†	0.0310

† not within the top five in this ranking
‡ not present in the original set of subsystems

not segregate from *Shigella* species, they tended to accumulate on one side of the two corresponding clusters, with Clades II to VIII gathering on the opposite side.

We can further group all seven clusters into three non-overlapping super-groups. These three super-clusters represent the evolutionary most distant strains according to a phylogenetic analysis (Fig 1b).

To identify the factors driving the observed clusters, we conducted an analysis of lPCA loadings grouped according to metabolic subsystems. Specifically, we computed average, subsystem-specific loading vectors as defined in (2), see Method section. Our investigation revealed that cluster separation was predominantly influenced by (the top) five subsystems: "Murein Biosynthesis", "Murein Recycling", "Anaplerotic Reactions", "Exchange", and "Histidine Metabolism" (Fig 1a). Notably, when employing a different ranking method based on (4), four out of these five driver subsystems remained consistent (see Tab 1), underscoring their robust impact on the clustering dynamics.

"Murein Recycling" and "Murein Biosynthesis" are known to be highly conserved subsystems among *Escherichia* strains, with 88 % of the reactions being shared across all *Escherichia* strains [8]. The highest loading values were obtained for D-alanyl-D-alanine dipeptidase ("ALAALAD" from the "Murein Recycling" subsystem), which is conserved in 82 % *E. coli*s, in 100 % *Shigella*s, and in 2 % Clades II to VIII. In contrast, "ALAALAD" is completely absent in the more distinct clades, *E. albertii* and *E. fergusonii*.

D-alanyl-D-alanine dipeptidase (*vanX*) is responsible for cleaving D-alanyl-D-alanine dipeptide. This enzyme allows bacteria to use D-alanine as a carbon source and modify peptidoglycan structures [36]. Intriguingly, modifications in these peptidoglycan precursors, such as the substitution of D-alanyl-D-alanine with D-alanyl-D-lactate, confer resistance to the antibiotic vancomycin, which targets terminal D-alanyl-D-alanine residues to prevent crosslinking [36, 37]. Given the potential absence of this enzyme in *E. albertii* and *E. fergusonii*, vancomycin might emerge as an effective therapeutic against infections caused by these bacteria.

In further exploring the capabilities of lPCA, we analyzed the set of differential reactions within the "Histidine Metabolism" subsystem when contrasting the clades *E. albertii* and *E. coli* (see supplementary Fig S1). We noted virtually identical loading values for the reactions "HISDr", "URCN", and "IZPN" (Tab 2), possibly suggesting a linear pathway. In fact, these reactions detail the stepwise breakdown of histidine into N-formimidoyl-L-glutamate via urocanate and 4-imidazolone-5-propanoate. This

Table 2. Reaction-specific loadings of the subsystem "Histidine metabolism"

Reaction	Loading PC1	Loading PC2
HISDr	-0.0043	0.0052
HIUH_c	0.0058	0.0027
IG3PS_1	0.0034	0.0172
IVCS	-0.0715	-0.0406
IZPN	-0.0045	0.0055
IZPN_1	-0.0006	0.0013
PRCOA1†	-0.0633	0.0441
PYROX_1	-0.0123	-0.0194
PYRZAM	-0.0162	0.1061
URCN	-0.0043	0.0052

† In the original models, PRCOA1 was associated with Histidine metabolism. We reassigned PRCOA1 the newly created metabolic subsystem "Cholesterol degradation". See text and Tab 1 for details.

pathway was found to be twice as prevalent in *E. albertii* strains (25 %) compared to *E. coli* strains (12 %). This differential trait suggests that at least certain *E. albertii* strains might possess the metabolic capability to utilize histidine as a source of carbon and nitrogen. This discernible biochemical distinction could aid in easily distinguishing *E. albertii* strains from pathogenic enteropathogenic *E. coli* (EPEC) and enterohemorrhagic *E. coli* (EHEC), as *E. albertii* strains are commonly mistaken for either EPEC or EHEC [38].

One particular reaction, "PRCOA1", prevalent in 61% of *E. albertii*-specific GSMMs, appeared as a significant difference, see Tab 2. However, we found that this reaction might be misallocated to "Histidine Metabolism". According to BIGG database [39], "PRCOA1" converts CoA-20-hydroxy-cholest-4-en-3-one C5 side chain to Androst-4-ene-3,17-dione, and belongs to "Cholesterol degradation", rather than "Histidine Metabolism". Thus, we newly introduced the so far missing subsystem "Cholesterol degradation", reassigned "PRCOA1" (and associated reactions according to KEGG), and repeated lPCA (see Fig 1a) as well as computing the subsystem loadings new (see Tab 1). This time "Cholesterol degradation" replaced "Histidine metabolism" among the top five driving subsystems.

Based on the results of Fig 1, we hypothesize that lPCA is able to discern phylogenetically distant species solely based on an analysis of the set of differential reactions in the corresponding GSMMs.

Comparison of 343 yeast-specific GSMMs

To test our hypothesis, we applied lPCA to the set of differential reactions extracted from 343 yeast-specific GSMMs [9], mostly obtained through a recent sequencing effort [27], see 2a. Similar to the lPCA analysis of GSMMs for *Escherichia* (see above), our findings reveal a clustering of species with closer phylogenetic relationships (compare Fig 2a and Fig 2b). However, when considering the first two principal components, the separation is less pronounced for yeast, accounting for only 27 % of the variance, in contrast to the above observed 34 % for *Escherichia* species. In the phylogenetic tree Fig 2b, similar clades were observed to group into closely related clusters in the lPCA plot Fig 2a. Notably, the *Lipomycetaceae* clade (purple) was distinctly isolated from other clades. In contrast, clades with closer phylogenetic relationships formed subclusters. In particular, the *Saccharomycodaceae*-specific GSMMs (green) constituted a distinct cluster, despite their close phylogenetic relation to *Saccharomycetaceae*.

Again, our analysis identified the top five subsystems critically contributing to

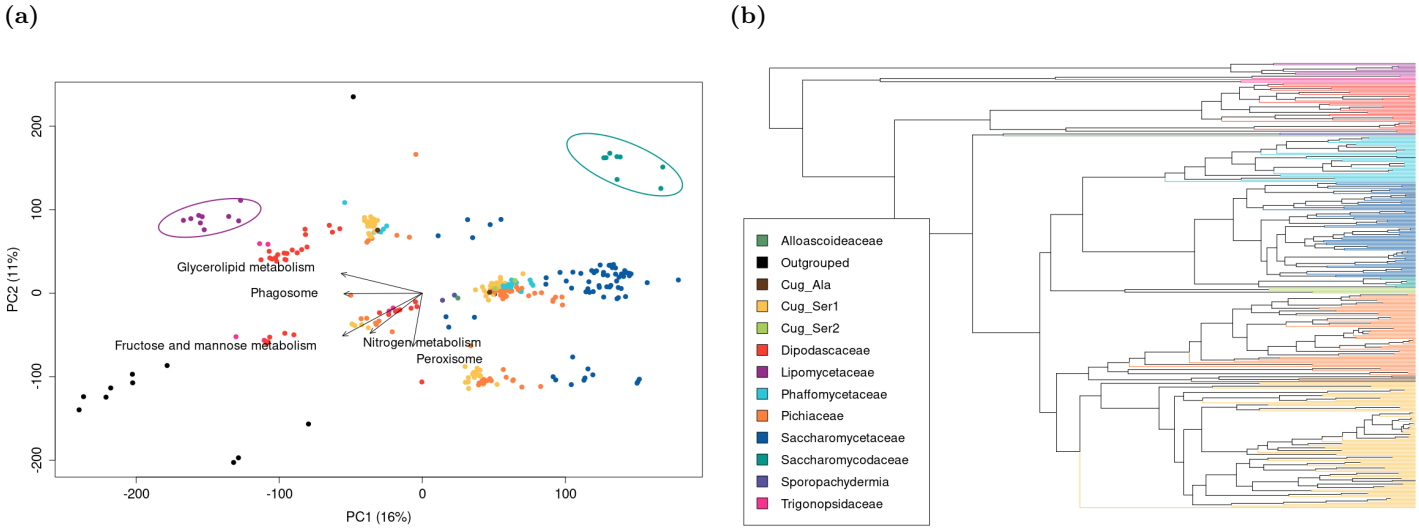


Fig 2. Relationship based on differential reactions across fungal GSMMs using lPCA compared to genome-based phylogeny. Related clades in the phylogenetic tree (b) tended to form closely related clusters in the lPCA plot (a). The most distinct clade (*Lipomycetaceae*) (purple) clustered sharply separated from other clades, while closer related clades clustered together in subclusters. *Saccharomycodaceae*-specific GSMMs (green) formed a clearly separated cluster in the lPCA, though phylogenetically they are closely related to *Saccharomycetaceae* (dark blue).

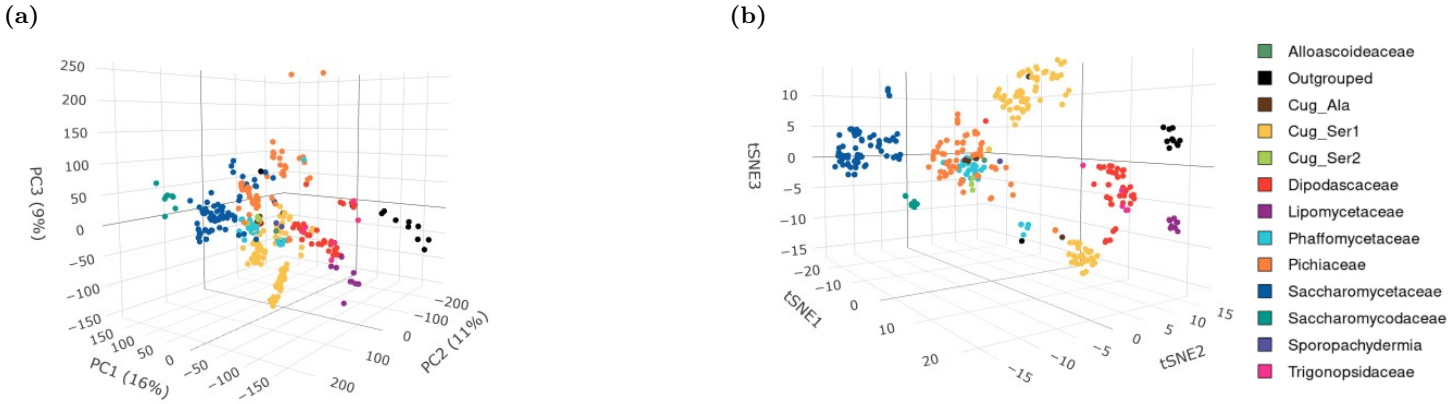


Fig 3. Comparison of three-dimensional (3D) clustering from lPCA and t-SNE using reaction matrices from GSMMs. Three principal components (a) from lPCA and three dimensions from t-SNE (b) were calculated for 343 fungal GSMMs. The resulting clustering is comparable across different clades between the two approaches. Both methods of visualization show that there are subclusters present in various clades within CUG-Ser1, and *Pichiaceae*, implying significant metabolic resemblances.

cluster separation: "Glycerolipid metabolism", "Phagosome", "Nitrogen metabolism", "Fructose and mannose metabolism", and "Peroxisome". However, comprehensively examining these pathways in yeast (as we did above for *Escherichia*) faces two major challenges (i) limited knowledge on fungal metabolism. In fact, yeast and other fungi remain relatively understudied compared to mammalian and prokaryotic cells [40]; and (ii) a significant proportion of unassigned reactions (>50 %) in the current data set.

Previously, Lu *et al.* [9] utilized a 3D t-SNE visualization of the binary matrix derived from the fungal GSMMs to identify similar groups of species. However,

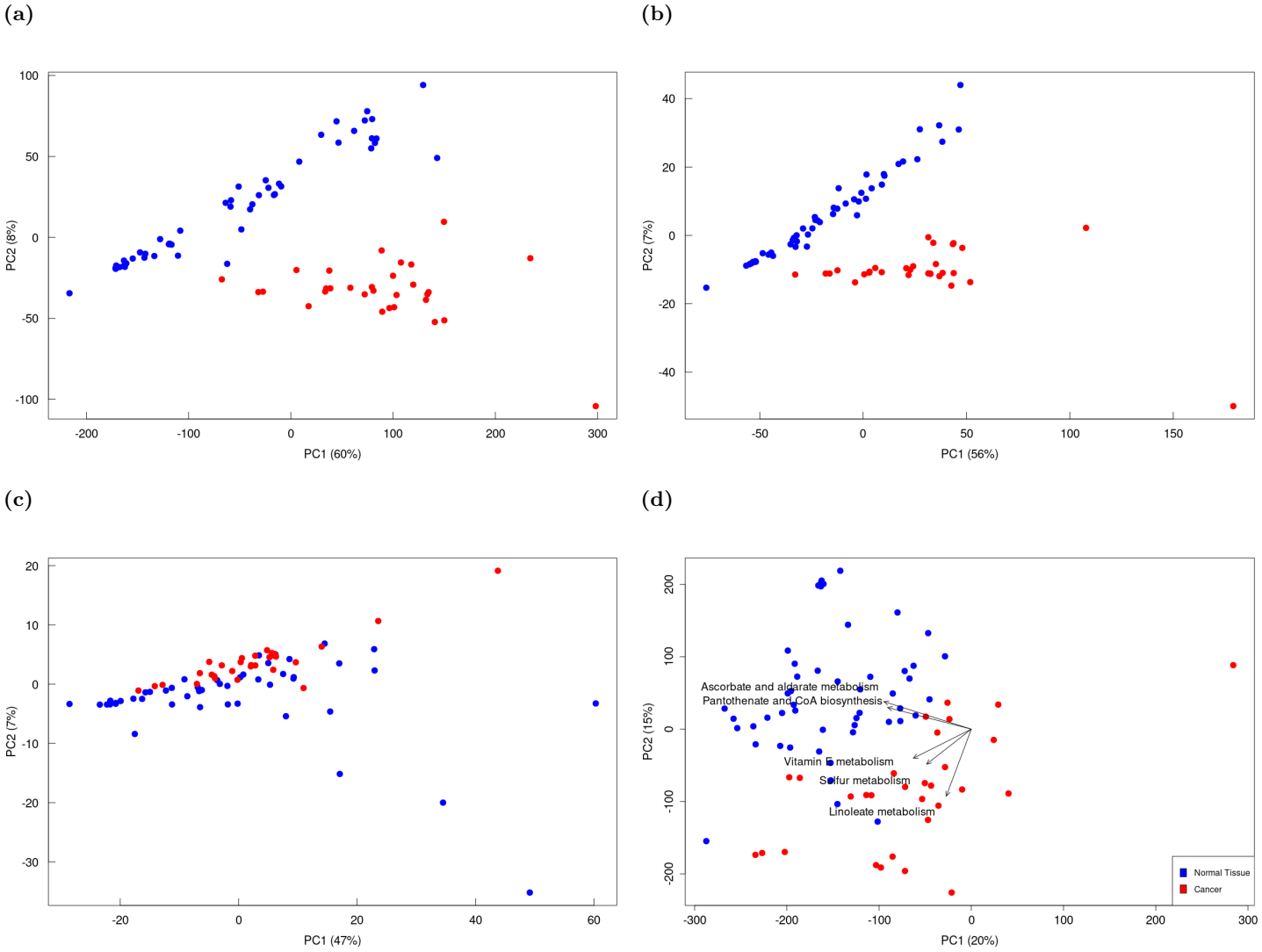


Fig 4. PCA and lPCA of transcriptomic and GSMM-derived datasets. PCA of transcriptomes from a) whole tissues, b) "Human1" metabolic genes, and c) selected genes from tissue-specific GSMMs. While clustering between healthy and cancer tissue could be conserved based on whole transcriptomes and metabolic genes, it was not possible based on the genes from the tissue-specific GSMMs. d) Separation based on lPCA of tissue-specific GSMMs from healthy and cancer tissues could be conserved.

employing 3D lPCA yields comparable results; compare Fig 3a and Tab 3b, and supplementary figures Fig S3 a and Tab S3 b.

Both visualization methods indicate subclusters across multiple clades in CUG-Ser1, and *Pichiaceae*, suggesting pronounced metabolic similarities. Our findings, however, reveal additional discernible subclusters within these clades.

However, certain clusters, particularly those from CUG-Ser1, *Phaffomycetaceae*, and *Pichiaceae*, had sharper delineation in the 3D t-SNE representation than in the 3D lPCA plot.

Comparison of healthy and cancerous tissues

Previous studies have demonstrated the effectiveness of PCA in distinguishing between healthy and cancerous tissues using transcriptomic data [41, 42]. Here, we replicated this finding using normalized gene expression data taken from the Human Protein Atlas [28, 29], see Fig 4a. When restricting the analysis to genes annotated within the generic "Human1" reconstruction, the distinction became less evident, see Fig 4b. Further narrowing the analysis to genes derived from the annotations in tissue-specific GSMMs (see Methods for details) resulted in the loss of differentiation (Fig 4c). This outcome may be attributed to the notably fewer genes (561) in these models compared to the broader set in "Human1" (containing 2897 metabolic genes). Consequently, we explored if lPCA could recover this differentiation using the differential set of reactions in these GSMMs. Indeed, Fig 4d indicates the feasibility of such recovery via lPCA.

For instance, in our analysis, "Linoleate Metabolism" emerged as a crucial subsystem for distinguishing between cancerous and healthy tissue-specific GSMMs. Within this subsystem, the reaction "MAR02438" exhibited the highest loading value and is associated with the genes *PTGS1* and *PTGS2*. These genes play a pivotal role in prostaglandin synthesis and are linked to the vascular endothelial-derived growth factor (VEGF) signaling pathway, critical for angiogenesis [43]. *PTGS2* holds significant importance in cancer research. Inhibitors targeting the enzyme COX2, encoded by the *PTGS2* gene, such as celecoxib, have demonstrated cancer-retarding properties [44].

Conclusion

Here we introduced logisitic principal component analysis (lPCA) to simultaneously compare and analyze multiple genome-scale metabolic models (GSMMs) to identify similarities and differences in metabolic capabilities across multiple species and strains. lPCA extends standard principal component analysis (PCA) to binary data sets. Thus, it allows us to analyze the presence or absence of biochemical reactions across GSMMs.

Our approach not only confirmed established phylogenetic relationships but also demonstrated the robustness of lPCA in delineating clustering patterns compared to traditional PCA.

Utilizing lPCA on the set of differential reactions across GSMMs provides distinct advantages:

1. Enhanced Clade Discrimination: Unlike PCA relying on simulated growth rates from varied environmental conditions, lPCA exhibited clear separation of distinct clades. This method mitigates bias by solely assessing reaction presence in GSMMs, rather than subjectively selecting environmental conditions for growth rate simulations.

2. Pathway Analysis and Subsystem Identification: In contrast to t-distributed stochastic neighborhood embedding (t-SNE) plots, our lPCA approach allowed precise information on drivers that govern separation and enables an efficient pathway analysis by grouping reactions based on their loading values. This may also support curation affords.

3. Transcriptome-based Equivalence: Our findings showcased the ability of lPCA to recover cluster patterns observed in transcriptome-based PCA plots solely from genome-scale metabolic reconstruction and/or GSMMs.

Acknowledgment

LZ, BK, and JAHB are employees of Baxalta Innovation GmbH. Employees of Baxalta Innovations GmbH may be owners of stock and/or stock options. DS and JZ are employees of the University of Vienna. The presented work was partly funded by Baxalta Innovations GmbH.

Data availability

All data and codes underlying this article are freely available in a GitHub repository, at `https://github.com/LeoZ93/lpca4gsmm`.

References

1. Schilling CH, Covert MW, Famili I, Church GM, Edwards JS, Palsson BO. Genome-scale metabolic model of Helicobacter pylori 26695; 2002.
2. Smith AC, Robinson AJ. A metabolic model of the mitochondrion and its use in modelling diseases of the tricarboxylic acid cycle. BMC systems biology. 2011;5(1):1–13.
3. Agren R, Mardinoglu A, Asplund A, Kampf C, Uhlen M, Nielsen J. Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. Molecular systems biology. 2014;10(3):721.
4. Mardinoglu A, Agren R, Kampf C, Asplund A, Uhlen M, Nielsen J. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nature communications. 2014;5(1):3083.
5. Ergün BG, Berrios J, Binay B, Fickers P. Recombinant protein production in Pichia pastoris: from transcriptionally redesigned strains to bioprocess optimization and metabolic modelling. FEMS Yeast Research. 2021;21(7):foab057.
6. Fouladiha H, Marashi SA, Torkashvand F, Mahboudi F, Lewis NE, Vaziri B. A metabolic network-based approach for developing feeding strategies for CHO cells to increase monoclonal antibody production. Bioprocess and Biosystems Engineering. 2020;43:1381–1389.
7. Richelle A, David B, Demaegd D, Dewerchin M, Kinet R, Morreale A, et al. Towards a widespread adoption of metabolic modeling tools in biopharmaceutical industry: a process systems biology engineering perspective. NPJ systems biology and applications. 2020;6(1):6.
8. Monk JM. Genome-scale metabolic network reconstructions of diverse Escherichia strains reveal strain-specific adaptations. Philosophical Transactions of the Royal Society B. 2022;377(1861):20210236.
9. Lu H, Li F, Yuan L, Domenzain I, Yu R, Wang H, et al. Yeast metabolic innovations emerged via expanded metabolic network and gene positive selection. Molecular Systems Biology. 2021;17(10):e10427.
10. Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, et al. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. Proceedings of the National Academy of Sciences. 2013;110(50):20338–20343.

11. Capela J, Lagoa D, Rodrigues R, Cunha E, Cruz F, Barbosa A, et al. merlin, an improved framework for the reconstruction of high-quality genome-scale metabolic models. Nucleic Acids Research. 2022;50(11):6052–6066.

12. Wang H, Marcišauskas S, Sánchez BJ, Domenzain I, Hermansson D, Agren R, et al. RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. PLoS computational biology. 2018;14(10):e1006541.

13. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic acids research. 2018;46(15):7542–7553.

14. Vlassis N, Pacheco MP, Sauter T. Fast reconstruction of compact context-specific metabolic network models. PLoS computational biology. 2014;10(1):e1003424.

15. Schultz A, Qutub AA. Reconstruction of tissue-specific metabolic networks using CORDA. PLoS computational biology. 2016;12(3):e1004808.

16. Zur H, Ruppin E, Shlomi T. iMAT: an integrative metabolic analysis tool. Bioinformatics. 2010;26(24):3140–3142.

17. Cabbia A, Hilbers PA, van Riel NA. A distance-based framework for the characterization of metabolic heterogeneity in large sets of genome-scale metabolic models. Patterns. 2020;1(6).

18. Heinken A, Hertel J, Acharya G, Ravcheev DA, Nyga M, Okpala OE, et al. Genome-scale metabolic reconstruction of 7,302 human microorganisms for personalized medicine. Nature Biotechnology. 2023; p. 1–12.

19. Lötsch J, Ultsch A. Current projection methods-induced biases at subgroup detection for machine-learning based data-analysis of biomedical data. International journal of molecular sciences. 2019;21(1):79.

20. Gove R, Cadalzo L, Leiby N, Singer JM, Zaitzeff A. New guidance for using t-SNE: Alternative defaults, hyperparameter selection automation, and comparative evaluation. Visual Informatics. 2022;6(2):87–97.

21. Ozgode Yigin B, Saygili G. Effect of distance measures on confidences of t-SNE embeddings and its implications on clustering for scRNA-seq data. Scientific Reports. 2023;13(1):6567.

22. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nature biotechnology. 2010;28(3):245–248.

23. Landgraf AJ, Lee Y. Dimensionality reduction for binary data through the projection of natural parameters. Journal of Multivariate Analysis. 2020;180:104668.

24. Collins M, Dasgupta S, Schapire RE. A generalization of principal components analysis to the exponential family. Advances in neural information processing systems. 2001;14.

25. Song Y, Westerhuis JA, Aben N, Michaut M, Wessels LF, Smilde AK. Principal component analysis of binary genomics data. Briefings in bioinformatics. 2019;20(1):317–329.

26. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: constraints-based reconstruction and analysis for python. BMC systems biology. 2013;7:1–6.

27. Shen XX, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, et al. Tempo and mode of genome evolution in the budding yeast subphylum. Cell. 2018;175(6):1533–1545.

28. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419.

29. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science. 2017;357(6352):eaan2507.

30. Robinson JL, Kocabaş P, Wang H, Cholley PE, Cook D, Nilsson A, et al. An atlas of human metabolism. Science signaling. 2020;13(624):eaaz1482.

31. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Achtman M, Brown D, et al. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. Genome research. 2020;30(1):138–152.

32. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics. 2010;11:1–11.

33. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome biology. 2019;20:1–14.

34. Paradis E, Blomberg S, Bolker B, Brown J, Claude J, Cuong HS, et al. Package 'ape'. Analyses of phylogenetics and evolution, version. 2019;2(4):47.

35. Krijthe J, van der Maaten L, Krijthe MJ. Package 'Rtsne'. R package version 013. 2018;.

36. Lessard IA, Walsh CT. VanX, a bacterial D-alanyl-D-alanine dipeptidase: resistance, immunity, or survival function? Proceedings of the National Academy of Sciences. 1999;96(20):11028–11032.

37. Aráoz R, Anhalt E, René L, Badet-Denisot MA, Courvalin P, Badet B. Mechanism-based inactivation of VanX, a D-alanyl-D-alanine dipeptidase necessary for vancomycin resistance. Biochemistry. 2000;39(51):15971–15979.

38. Gomes TA, Ooka T, Hernandes RT, Yamamoto D, Hayashi T. Escherichia albertii pathogenesis. EcoSal Plus. 2020;9(1):10–1128.

39. Norsigian CJ, Pusarla N, McConn JL, Yurkovich JT, Dräger A, Palsson BO, et al. BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. Nucleic acids research. 2020;48(D1):D402–D406.

40. Roth M, Westrick N, Baldwin T. Fungal biotechnology: From yesterday to tomorrow. Front Fungal Biol. 2023; 4; 2023.

41. Shaath H, Toor S, Nair VS, Elkord E, Alajez NM. Transcriptomic analyses revealed systemic alterations in gene expression in circulation and tumor microenvironment of colorectal cancer patients. Cancers. 2019;11(12):1994.

42. Xu L, Wang R, Ziegelbauer J, Wu WW, Shen RF, Juhl H, et al. Transcriptome analysis of human colorectal cancer biopsies reveals extensive expression correlations among genes related to cell proliferation, lipid metabolism, immune response and collagen catabolism. Oncotarget. 2017;8(43):74703.

43. Kamal MV, Damerla RR, Dikhit PS, Kumar NA. Prostaglandin-endoperoxide synthase 2 (PTGS2) gene expression and its association with genes regulating the VEGF signaling pathway in head and neck squamous cell carcinoma. Journal of Oral Biology and Craniofacial Research. 2023;13(5):567–574.

44. Xu L, Stevens J, Hilton MB, Seaman S, Conrads TP, Veenstra TD, et al. COX-2 inhibition potentiates antiangiogenic cancer therapy and prevents metastasis in preclinical models. Science translational medicine. 2014;6(242):242ra84–242ra84.

Supplementary Information

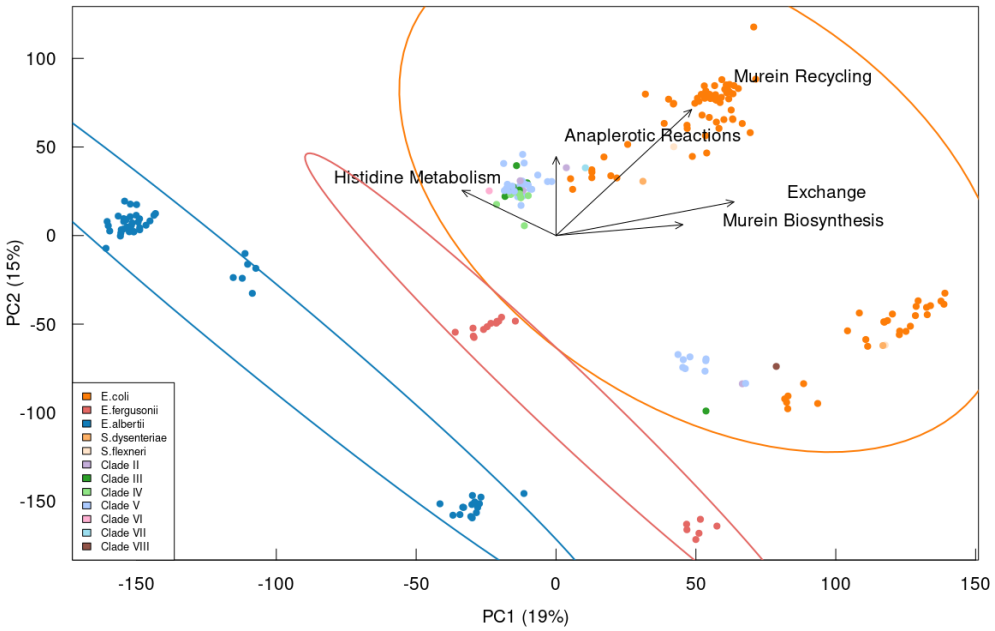


Fig S1 . Logisitic principal component analysis (lPCA) of *Escherichia* strains with original reaction assignments to subsystems. Before reassignment of "PRCOA1" from "Histidine Metabolism" to the newly created subsystem "Choleterol degradation", "Histidine Metabolism" was among the five pivotal subsystems driving the separation. However, this was only due to the wrongly assigned reaction "PRCOA1".

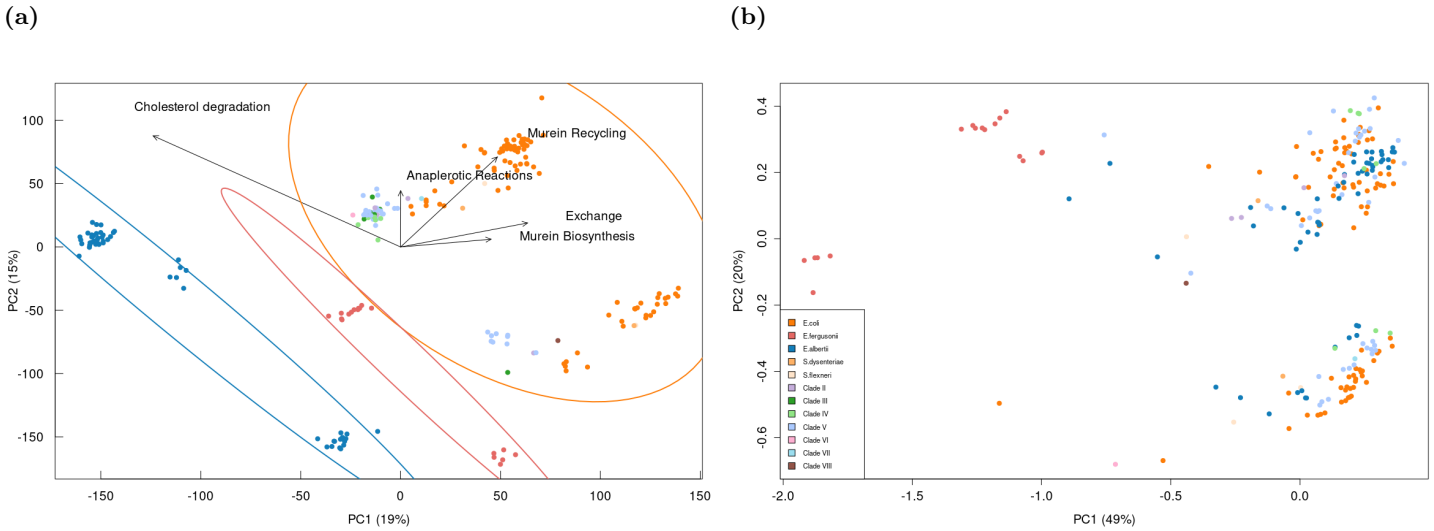


Fig S2 . Principal component analysis (PCA) and lPCA of 222 *Escherichia* strains (a) LPCA based on differential reactions across genome-scale metabolic models (GSMMs). (b) PCA based on simulated growth rates from 222 *Escherichia* species-specific GSMMs across 570 different environmental conditions [8]. The plot was reproduced from Monk [8] and resulted in similar clustering as in the original paper. *E. fergusonii* was the most distinct clade, while the remaining clades tended to form subclusters, instead of clade-specific clusters.

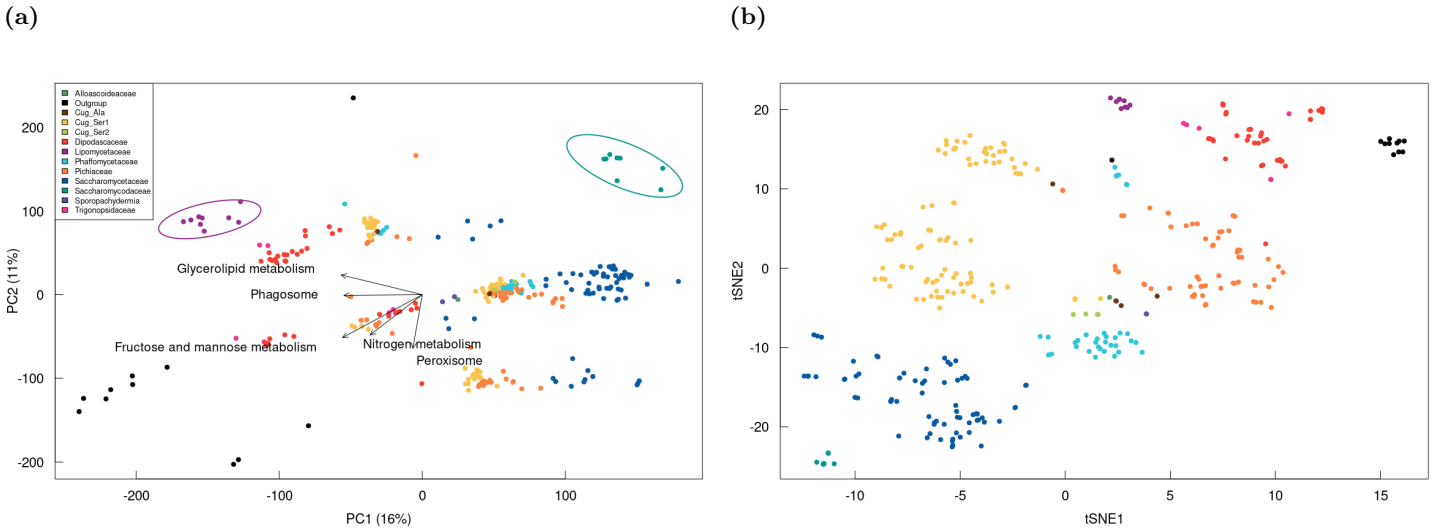


Fig S3 . Comparison of 2d clustering from lPCA and t-distributed stochastic neighborhood embedding (t-SNE) using reaction matrices from GSMMs. Two principal components (a) from lPCA and two dimensions from t-SNE (b) were calculated for 343 fungal GSMMs. The resulted clustering is comparable across different clades between the two approaches. Both methods of visualization show that there are subclusters present in various clades within CUG-Ser1, *Pichiaceae*, and *Phaffomycetaceae*, implying significant metabolic resemblances. Our research uncovers additional distinguishable subclusters within these clades, especially in *Pichiaceae*, and CUG-Ser1.