**Links between gut microbiome composition and fatty liver disease in a large population sample**

**Supplementary figures and tables**

Matti O. Ruuskanen1,2\*, Fredrik Åberg3,4, Ville Männistö5,6, Aki S. Havulinna2,7, Guillaume Méric8,9, Yang Liu8,10, Rohit Loomba11,12, Yoshiki Vázquez-Baeza13,14, Anupriya Tripathi15,16,17, Liisa M. Valsta2, Michael Inouye8,18, Pekka Jousilahti2, Veikko Salomaa2, Mohit Jain12,19, Rob Knight13,14,20,21, Leo Lahti22, Teemu J. Niiranen1,2,23

1Department of Internal Medicine, University of Turku, Turku, Finland   
2Department of Public Health Solutions, Finnish Institute for Health and Welfare, Helsinki, Finland  
3Transplantation and Liver Surgery Clinic, Helsinki University Hospital, University of Helsinki, Helsinki, Finland  
4The Transplant Institute, Sahlgrenska University Hospital, Gothenburg, Sweden  
5Department of Medicine, Kuopio University Hospital, University of Eastern Finland, Kuopio, Finland  
6Department of Experimental Vascular Medicine, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands  
7Institute for Molecular Medicine Finland, FIMM - HiLIFE, Helsinki, Finland.  
8Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia

9Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria, Australia

10Department of Clinical Pathology, The University of Melbourne, Melbourne, Victoria, Australia

11Department of Medicine, NAFLD Research Center, La Jolla, CA, USA

12Department of Medicine, University of California, San Diego, La Jolla, CA, USA

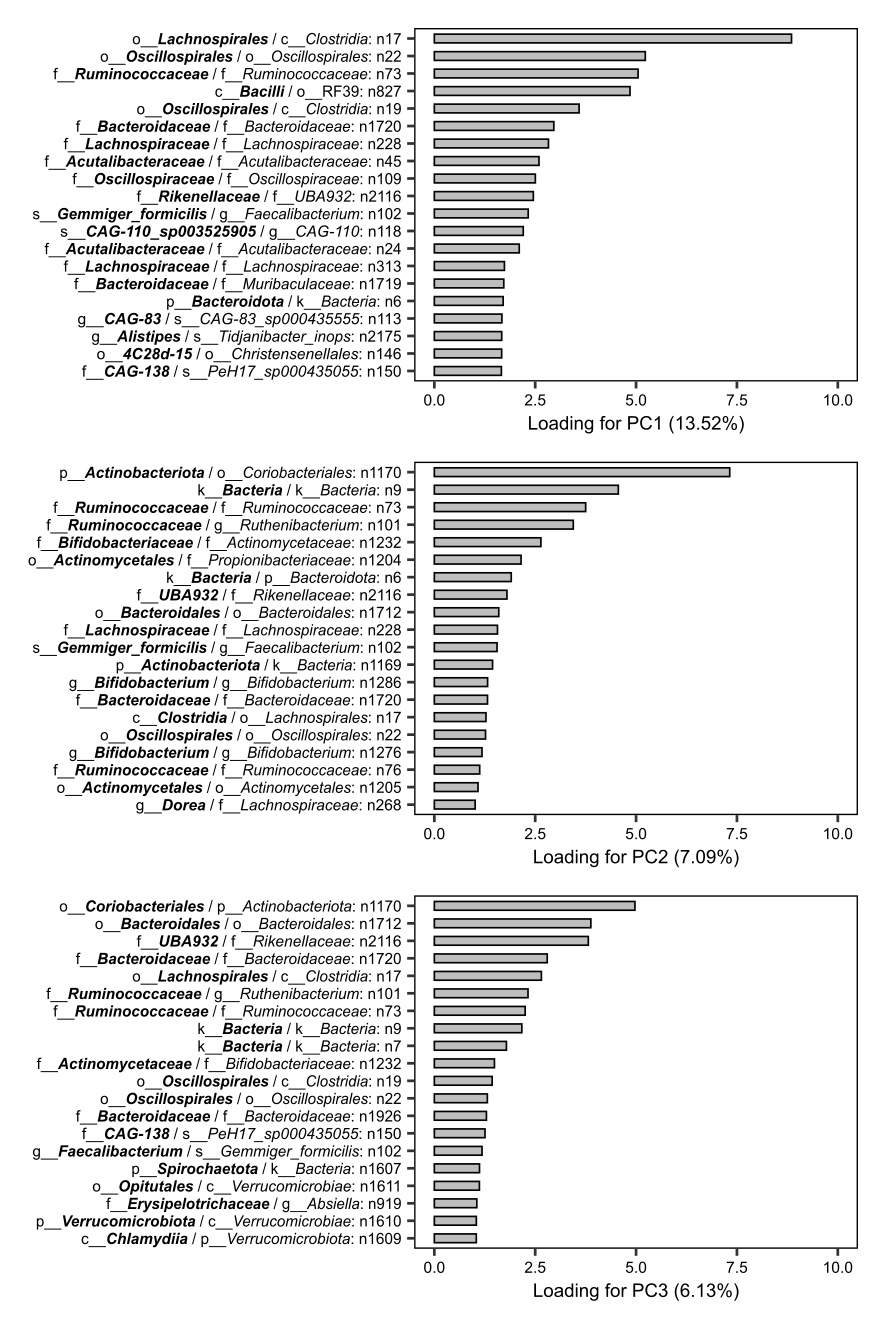
13Jacobs School of Engineering, University of California, San Diego, La Jolla, CA, USA

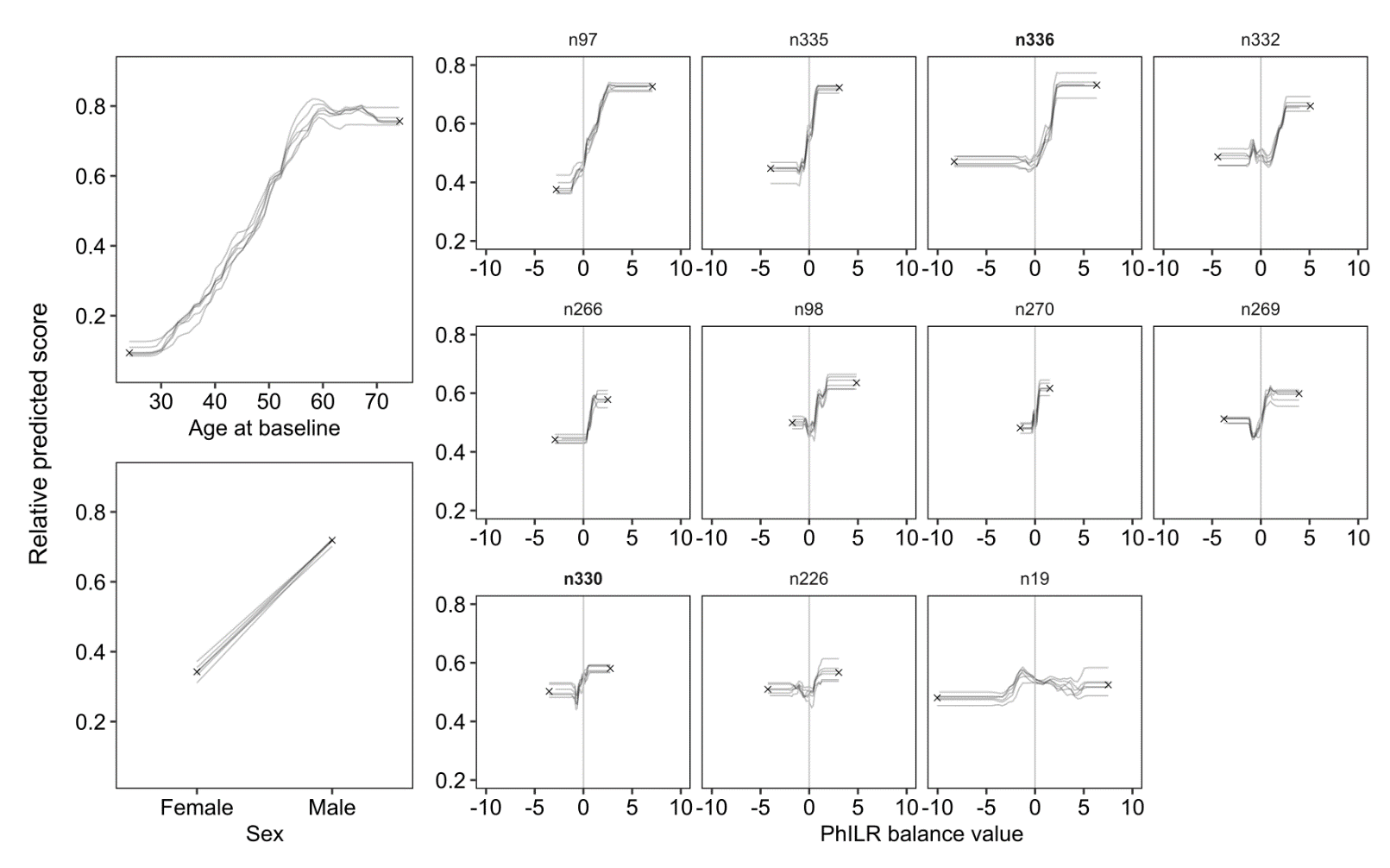
14Center for Microbiome Innovation, University of California San Diego, La Jolla, California, USA

15Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, California, USA

16Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, California, USA

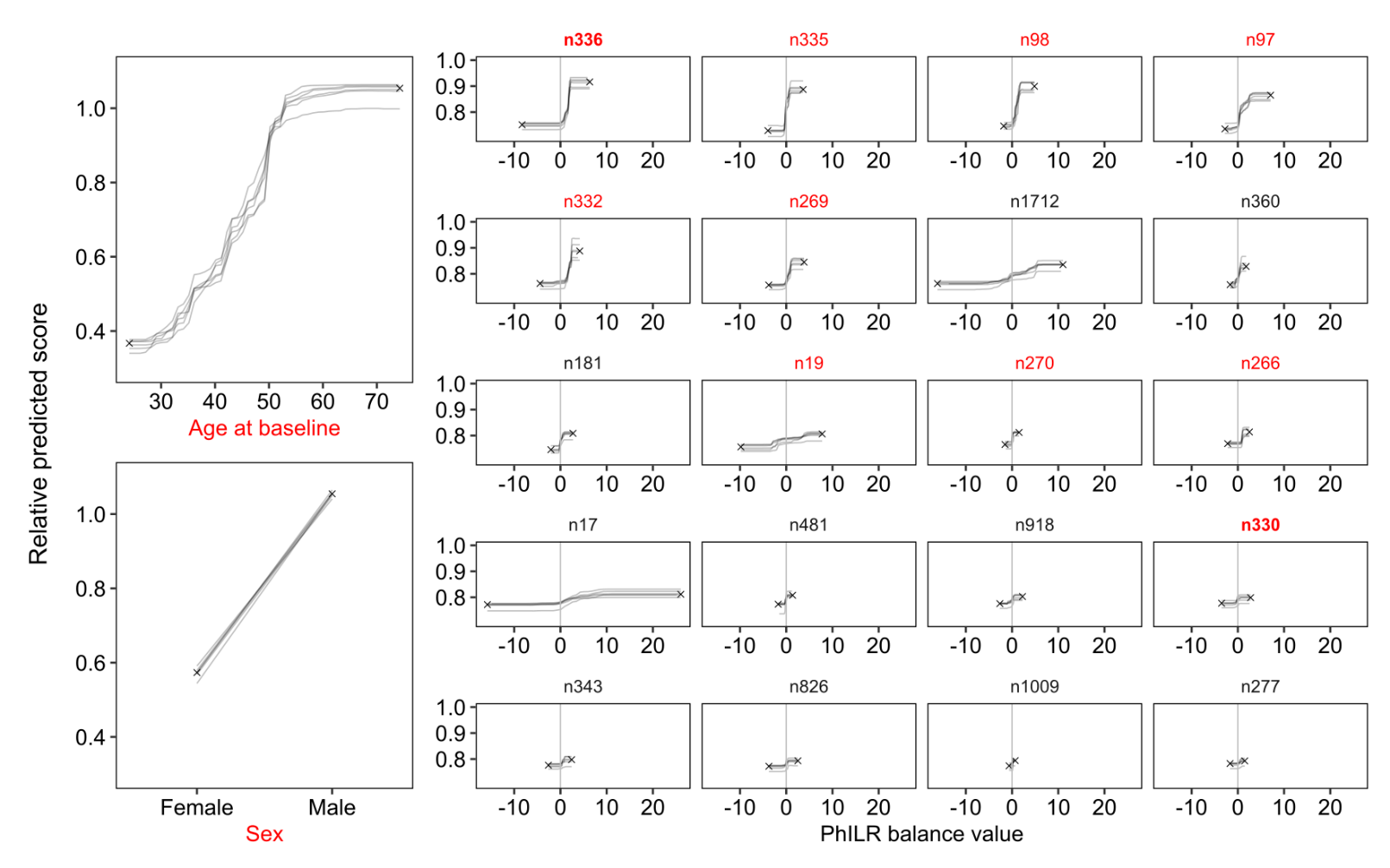
17Division of Biological Sciences, University of California, San Diego, La Jolla, California, USA  
18Department of Public Health and Primary Care, Cambridge University, Cambridge, United Kingdom  
19Department of Pharmacology, University of California San Diego, La Jolla, California, USA  
20Department of Pediatrics, School of Medicine, University of California San Diego, La Jolla, California, USA  
21Department of Computer Science & Engineering, University of California San Diego, La Jolla, California, USA  
22Department of Future Technologies, University of Turku, Turku, Finland  
23Division of Medicine, Turku University Hospital, Turku, Finland

Figure S1. Loadings of the top 20 balances in the bacterial phylogenetic tree for the first principal components. Taxonomic information is based on the GTDB standardized taxonomy (Parks et al., 2018) and the taxa are named based on a simple voting scheme of the two descendent clades of a given balance separately (in PhILR; Silverman et al., 2017). Balances have been rotated and the signs of their loadings on the PC axes have been switched accordingly, so that the indicated sides of the balances have a positive association with FLI.

Figure S2. Partial dependence of features on the prediction outcome in the classification models discerning between the FLI < 60 and FLI ≥ 60 groups. Relative predicted score indicates the probability of classification in the FLI ≥ 60 group calculated independently of other features. Individual lines for the features show the variability of the partial dependence in the 6 LOGOCV models. The medians of the relative predicted scores at the minimum and maximum values of each feature (used in Figure 2) are shown as black crosses. The liver-specific balances are indicated with bold font. Because the signs of the PhILR balance values are dependent on the orientation of the branches at each balance (and these can be freely rotated in a phylogenetic tree), all partial dependences of the balances are shown here in the positive direction. The changes in relative predicted scores for each feature and the specific clades under each balance in this figure are also summarized in Figure 2.

**<Because of its large size, Figure S3 is supplied as an external file to retain legibility: Figure\_S3.eps>**

**Figure S3**. Complete phylogenetic tree of taxa covered by the 11 predictive balances of the FLI classification model. Relative predicted scores in the models, number of pathways for relevant processes and presence of specific acetate and ethanol production pathways in the representative genomes are included. The maximum number of pathways is different per class of pathway. Branch distance in this cladogram is not proportional to evolutionary distance. The figure is based on the GTDB standardized bacterial taxonomy (Parks et al., 2018), and includes species and other taxa not found in the NCBI taxonomy. The liver-specific balances are indicated with bold font.

Figure S4. Partial dependence of features on the prediction outcome in the classification models discerning between FLI < 30 and FLI ≥ 60 groups. In these models, participants with FLI ≥ 30 but < 60 have been excluded. Relative predicted score indicates the probability of classification in the FLI ≥ 60 group calculated independently of other features. Individual lines for the features show the variability of the partial dependence in the 6 LOGOCV models. The medians of the relative predicted scores at the minimum and maximum values of each feature are shown as black crosses. Features in these models common with the main model are indicated with red font (only balance n226 was not included in the selected features). The liver-specific balances are indicated with bold font. Because the signs of the PhILR balance values are dependent on the orientation of the branches at each balance (and these can be freely rotated in a phylogenetic tree), all partial dependences of the balances are shown here in the positive direction.

**Table S1.** Classification model validation results with AUC. AUC are shown for all classification models and summarized per region omitted in training of the model but used for its validation. The number of participants in these validation sets are indicated in bold, and their units differ from other columns.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Main model (FLI < 60 / FLI ≥ 60) | | | | Alternative model (FLI < 30 / FLI ≥ 60) | | | |
| CV region | Full model | Covariates only | Balances  only | *N*  participants | Full model | Covariates only | Balances  only | *N*  participants |
| North Karelia | 0.761 | 0.712 | 0.671 | **379** | 0.801 | 0.762 | 0.719 | **282** |
| Lapland | 0.734 | 0.664 | 0.676 | **389** | 0.772 | 0.722 | 0.715 | **288** |
| North Savonia | 0.788 | 0.740 | 0.681 | **288** | 0.800 | 0.750 | 0.673 | **213** |
| Oulu province | 0.730 | 0.661 | 0.669 | **249** | 0.773 | 0.721 | 0.720 | **190** |
| Turku/Loimaa | 0.755 | 0.760 | 0.598 | **245** | 0.784 | 0.781 | 0.630 | **186** |
| Helsinki/Vantaa | 0.743 | 0.731 | 0.647 | **326** | 0.891 | 0.839 | 0.753 | **242** |
| Mean | 0.752 | 0.711 | 0.657 | **312.7** | 0.804 | 0.763 | 0.702 | **233.5** |

**Table S2.** Classification model validation results with AUPRC. AUPRC are shown for all classification models and summarized per region omitted in training of the model but used for its validation. The number of participants in the validation sets are indicated in bold, and the proportion of them in the FLI ≥ 60 group are indicated in italic.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Main model (FLI < 60 / FLI ≥ 60) | | | | | Alternative model (FLI < 30 / FLI ≥ 60) | | | | |
| CV region | Full model | Covariates only | Balances  only | *N*  participants | Proportion of  FLI ≥ 60 | Full model | Covariates only | Balances  only | *N*  participants | Proportion of  FLI ≥ 60 |
| North Karelia | 0.602 | 0.494 | 0.496 | **379** | *0.311* | 0.765 | 0.717 | 0.686 | **282** | *0.440* |
| Lapland | 0.576 | 0.456 | 0.470 | **389** | *0.314* | 0.737 | 0.643 | 0.642 | **288** | *0.424* |
| North Savonia | 0.597 | 0.469 | 0.500 | **288** | *0.285* | 0.770 | 0.680 | 0.646 | **213** | *0.446* |
| Oulu province | 0.524 | 0.385 | 0.534 | **249** | *0.273* | 0.695 | 0.570 | 0.651 | **190** | *0.384* |
| Turku/Loimaa | 0.533 | 0.522 | 0.400 | **245** | *0.298* | 0.729 | 0.739 | 0.524 | **186** | *0.366* |
| Helsinki/Vantaa | 0.553 | 0.498 | 0.430 | **326** | *0.304* | 0.813 | 0.751 | 0.646 | **242** | *0.376* |
| Mean | 0.564 | 0.471 | 0.472 | **312.7** | *0.297* | 0.752 | 0.683 | 0.633 | **233.5** | *0.406* |

**Table S3.** Performance of classification models discerning between FLI ≥ 60 and FLI < 60 groups, with data from the validation region excluded from feature selection and hyperparameter optimization. The number of participants in the validation sets are indicated in bold, and the proportion of them in the FLI ≥ 60 group are indicated in italic.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | AUC | | | | AUPRC | | |  | |
| CV region | Full model | Covariates only | Balances  only | Full model | | Covariates only | Balances  only | *N*  participants | Proportion of  FLI ≥ 60 |
| North Karelia | 0.755 | 0.712 | 0.670 | 0.602 | | 0.493 | 0.485 | **379** | *0.311* |
| Lapland | 0.742 | 0.670 | 0.700 | 0.585 | | 0.495 | 0.504 | **389** | *0.314* |
| North Savonia | 0.791 | 0.729 | 0.686 | 0.611 | | 0.458 | 0.494 | **288** | *0.285* |
| Oulu province | 0.737 | 0.656 | 0.676 | 0.543 | | 0.375 | 0.536 | **249** | *0.273* |
| Turku/Loimaa | 0.737 | 0.755 | 0.590 | 0.531 | | 0.495 | 0.407 | **245** | *0.298* |
| Helsinki/Vantaa | 0.749 | 0.729 | 0.668 | 0.571 | | 0.488 | 0.448 | **326** | *0.304* |
| Mean | 0.752 | 0.709 | 0.665 | 0.574 | | 0.467 | 0.479 | **312,7** | *0.297* |

**Table S4**. MetaCyc pathways analyzed in the representative genomes of GTDB taxa for production of ethanol and short chain fatty acids (acetate, butyrate, propionate), bile acid metabolism, and choline degradation to trimethylamine. One pathway might be included under multiple processes, based on its classification in MetaCyc.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Process classification | MetaCyC pathway | Main substrate(s) | Product(s) | In *N* genomes |
| Fermentation to Acetate | P41-PWY | pyruvate | **acetate**, (S)-lactate | **-** |
|  | PWY-5096 | pyruvate | **acetate**, L-alanine | **-** |
|  | PWY-5100 | pyruvate | **acetate**, (S)-lactate, CO2 | **-** |
|  | P142-PWY | pyruvate | **acetate**,CO2 | **-** |
|  | PWY-5482 | pyruvate | **acetate** | 7 |
|  | PWY-5483 | pyruvate | **acetate** | - |
|  | PWY-5485 | pyruvate | **acetate**, formate | - |
|  | PWY-5537 | pyruvate | **acetate**, succinate | - |
|  | PWY-5538 | pyruvate | **acetate**, succinate, CO2 | - |
|  | PWY-5600 | pyruvate | **acetate**,CO2 | - |
|  | PWY-5768 | pyruvate | **acetate** | - |
|  | PWY0-1312 | acetyl-CoA | **acetate** | 204 |
|  | PWY-5535 | acetyl-CoA | **acetate** | - |
|  | PWY-5536 | acetyl-CoA | **acetate** | - |
|  | P161-PWY | acetylene | **acetate**, ethanol | 10\* |
|  | P124-PWY | D-glucopyranose | **acetate**, (S)-lactate | - |
|  | P461-PWY | hexitols | **acetate**, ethanol, (S)-lactate | - |
|  | PROPFERM-PWY | L-alanine | **acetate**, propanoate, ammonium, CO2 | - |
|  | P162-PWY | L-glutamate | **acetate**, butanoate | 42\* |
|  | P163-PWY | L-lysine | **acetate**, butanoate | 12\* |
| Fermentation to Butanoate | CENTFERM-PWY | pyruvate | **butanoate** | 34 |
|  | PWY-5677 | succinate | **butanoate** | 3 |
|  | PWY-5022 | 4-aminobutanoate | **butanoate**, acetate, ammonium | 8 |
|  | PWY-5676 | acetyl-CoA | **butanoate** | - |
|  | P162-PWY | L-glutamate | **butanoate**, acetate | 42\* |
|  | GLUDEG-II-PWY | L-glutamate | **butanoate**, H2 | - |
|  | P163-PWY | L-lysine | **butanoate**, acetate | 12\* |
| Fermentation to Propanoate | P108-PWY | pyruvate | **propanoate** | 7 |
|  | PWY-5494 | (R)-lactate | **propanoate** | - |
|  | PWY-8086 | (S)-lactate | **propanoate**, acetate, H2 | 1 |
|  | PWY-7013 | propylene glycol | **propanoate**, propan-1-ol | 6 |
|  | PROPFERM-PWY | L-alanine | **propanoate**, acetate, ammonium, CO2 | - |
|  | PWY-5088 | L-glutamate | **propanoate** | - |
| Fermentation to Ethanol | PWY-5480 | pyruvate | **ethanol**, formate | 68 |
|  | PWY-5486 | pyruvate | **ethanol** | - |
|  | PWY-6587 | pyruvate | **ethanol** | 25 |
|  | FERMENTATION-PWY | phosphoenolpyruvate | **ethanol**, acetate, succinate, 2-oxoglutarate, (R)-lactate, CO2, H2 | 6 |
|  | P161-PWY | acetylene | **ethanol**, acetate | 10\* |
|  | P122-PWY | β-D-fructofuranose, D-glucopyranose | **ethanol**, (S/R)-lactate, CO2 | 34 |
|  | P461-PWY | hexitols | **ethanol**, acetate, (S)-lactate | - |
| Bile Acid Degradation | PWY-7754 | cholate / chenodeoxycholate | **deoxycholate /  lithocholate** | 98 |
|  | PWY-6518 | cholate | **7-epicholate**, **12-epicholate, isocholate** | 8 |
| Choline Degradation to TMA | PWY-7167 | choline | **trimethylamine** | 24 |
|  | PWY-3641 | L-carnitine | **trimethylamine**, pyruvate | - |

\*Count duplicated, because pathway is included under more than one process classification.