PNAS
www.pnas.org

Supporting Information for

Gut microbiome contributions to altered metabolism in a pig model of undernutrition

Hao-Wei Chang, Nathan P. McNulty, Matthew C. Hibberd, David O'Donnell, Jiye Cheng, Vincent Lombard, Bernard Henrissat, Olga Ilkayeva, Michael J. Muehlbauer, Christopher B. Newgard, Michael J. Barratt, Xi Lin, Jack Odle, and Jeffrey I. Gordon

Corresponding author: Jeffrey Gordon
email: jgordon@wustl.edu

This PDF file includes:

Supplementary Results
Figures S1-S7
SI References

Other supplementary materials include:

Datasets S1 to S7

SUPPLEMENTARY RESULTS

Entropy-based method for microbial ecology research (EMMER)

Mathematical framework – A mathematical procedure was needed for transferring a non-density matrix composed of CAZyme (or ASV) features to a density matrix that satisfies the requirements for calculating the vNE. For a density matrix *D* that has a set of eigenvalues $\lambda = \{\lambda_1, \lambda, \ldots, \lambda_n\}$, the vNE of *D* can be calculated as follows:

$$H = -\sum_{i}^{n} \lambda_i \log_2 \lambda_i \quad [1]$$

Let matrix *A* be a non-density matrix that contains only real numbers, and matrix *C* be the covariance matrix of *A*. *C* satisfies the first two properties of a density matrix (a symmetric, positive semi-definite matrix; *1*). Because the sum of eigenvalues equals the trace of a matrix (*1*), we can satisfy the third mathematical requirement of a density matrix (trace of one) by normalizing the eigenvalue by the sum of all eigenvalues (Fig. S2*A*). Let $\alpha = \{\alpha_1, \alpha, \ldots, \alpha_n\}$ be the set of eigenvalues of *C*:

$$\beta_i = \frac{\alpha_i}{\sum \alpha} \quad [2]$$

We used normalized eigenvalues (β) from [2] to calculate the vNE:

$$H = -\sum_{i}^{n} \beta_i \log_2 \beta_i \qquad [3]$$

Feature selection and the reproducibility of information-rich feature calling – To evaluate the information content of each feature (column) in a non-density matrix, we calculated the vNE after systematically removing one column from the matrix at a time and generating a set of the vNE $\{H_1, H_2, \ldots, H_n\}$. For example, we generated matrix A_p by removing the p^{th} column from A, then followed the mathematical procedure described above to calculate the vNE (H_p). Removing a feature containing important information in the matrix will result in a marked change in vNE. A marked change in vNE is defined as the vNE above an upper or below a lower selection threshold (Fig. S2*B*): For initial analysis, thresholds were set at ± 2 standard deviations of values in $\{H_1, H_2, \ldots, H_n\}$. Using our previous example H_p, if H_p is outside the selection thresholds, then the corresponding feature of the p^{th} column in A was designated an "information-rich feature candidate". If an information-rich feature candidate was nominated at least twice during jackknife resampling, that candidate was included in the final list of information-rich features (Fig. S2*C*). We calculated the reproducibility of information-rich feature calling by dividing the number of times that a specific feature was nominated as an information-rich feature candidate by the number of jackknife resampling runs.

Threshold selection – Because of the shared procedure used for PCA and our mathematical framework, we reasoned that a PCA plot generated from information-rich features will retain the original data distribution generated from the full matrix in PCA space. Procrustes score *s* represents the similarity between the projections of the original input matrix prior to feature selection and an input matrix composed of information-rich features in PCA space. Procrustes score *w* represents the similarity between projections of the original input matrix and a matrix that only contains non-information-rich features in PCA space. When EMMER selects features that recapitulate the projection of original input data in PCA space, we expect the dissimilarity score, calculated by dividing *s* by *w*, to be low. We refined the thresholds for selecting information-rich features by first describing the linear relationship between dissimilarity score and the number of information-rich features. We then chose a threshold corresponding to a dissimilarity score that had the greatest distance below the regression line (Fig. S2*D*). In practice, we used the standard deviation of $\{H_1, H_2, \ldots, H_n\}$ as the unit for feature-calling thresholds. The feature-calling threshold was subsequently optimized by testing all combinations of upper and lower thresholds from 1.5 to 2.5 units at increments of 0.25 unit (a unit corresponds to one standard deviation of $\{H_1, H_2, \ldots, H_n\}$) (Dataset S2).

Improved computational accuracy – Calculating a covariance matrix can generate extremely small values; these small values can, in turn, decrease computational accuracy. After singular value decomposition (SVD), the square of singular values from an input matrix is equal to eigenvalues of a covariance matrix of that input matrix (*1*). Therefore, applying SVD in EMMER allows us to circumvent the covariance matrix generating step without changing the result of the vNE calculation.

Identification of information-rich features that differentiate treatment groups in PCA space – Applying SVD also allowed us to identify information-rich features that differentiate treatment groups in PCA space. For given matrix B where rows represent animals and columns represent information-rich microbial community features, the relationship between animals and specific principal components (PCs) is separated into a set of matrices $\{B_1, B_2, \ldots, B_k\}$ after SVD (Fig. S3*A*). Each matrix can be expressed as:

$$B_i = U \times D_i \times V^T \quad [4]$$

where B_i preserves the relationship between animals and the i^{th} PC. U and V^T are the left and right singular matrix, respectively. All elements in D_i are zero except for element $d_{i,i}$, which contains the i^{th} singular value.

In the current study, we visualized our data in PCA plots that contained the first three principal components (PCs). Therefore, we sought to identify information-rich features that differentiate FF and DR in the first three PCs. Fig. S3*A* and equation [5] describe our method for cleaning the matrix by removing information beyond PC3:

$$B_{III} = U \times (D_1 + D_2 + D_3) \times V^T \quad [5]$$

After cleaning the matrix, a linear regression model is built from B_{III} by using treatment groups as the response and information-rich features as explanatory variables. Differentiating information-rich features are explanatory variables that pass a significant threshold. Fifty of the 98 information-rich CAZymes and 40 of the 59 information-rich ASVs identified as differentiating FF and DR fecal microbial communities in PCA space are shown in Fig. S3*B*,*C*.

Evaluating the EMMER algorithm

We applied EMMER to a well-characterized, previously published V4-16S rDNA amplicon sequencing dataset (*2*) to demonstrate that it is a flexible feature selection algorithm. The dataset describes the bacterial composition of fecal microbiota sampled from adult lean human subjects practicing chronic calorie restriction with adequate nutrition (CRON; n=34) and from subjects consuming a typical unrestricted USA diet (AMER; n=66). Using Random Forests, indicator species and phi-correlation analyses, Griffin et al. (*2*) identified 242 dietary practice (DP)-associated bacterial taxa (operational taxonomic units, OTUs) that differentiate AMER and CRON microbiota in Principal Coordinates Analysis (PCoA; unweighted UniFrac distances; Fig S4*A*; Dataset S6). We used this dataset to demonstrate the flexibility of EMMER by exploring two analytical scenarios, one in which the groups considered are highly distinct, and another in which they are highly similar.

Two sample groups with dissimilar microbiota configurations - We first tested whether the features selected by EMMER capture enough information from the underlying data to differentiate two relatively distinct groups. Feature selection was performed on the AMER and CRON datasets individually using EMMER to identify group-specific information-rich taxa, after which the feature sets were combined to create an aggregate set of 72 information-rich taxa (Dataset S6). Using PCoA, we visualized unweighted UniFrac distances between communities based on the representation of the originally described 242 DP-associated taxa (Fig. S4*A*; Dataset S6) or the 72 information-rich taxa identified by EMMER (Fig. S4*B*; Dataset S6). In both cases, the distribution of CRON samples is statistically different from AMER samples in PCoA space ($P<0.005$; PERMANOVA), indicating information-rich taxa can be used to distinguish the two sample groups.

Two sample groups with similar microbiota configurations - We next tested whether the features identified by EMMER are consistent when the underlying data come from highly similar groups. Therefore, data from the 34 CRON subjects were randomly allocated to one of two groups (n=17/subset). In total, 31 information-rich taxa were nominated by EMMER; 26 of the information-rich taxa were shared between the two subsets. The distribution of the two CRON subsets exhibited no significant difference by PCoA ($P>0.05$; PERMANOVA; Fig. S5*A*; Dataset S6). To assess the reproducibility of this finding, we repeated the previous steps, including random sub-setting of the CRON dataset, EMMER analysis, and PCoA visualization, for nine additional iterations. A total of 82% ± 9% (mean ± SD) of all information-rich taxa identified in both subsets in a given iteration were shared (Fig. S5*B*). In all cases, the difference in distributions of each pair of subsets did not achieve statistical significance in PCoA space (Fig. S5*C*).

Assessing the effect of litter membership on pig gut microbiota configuration

Fig. S6 compares the pig gut microbiota in each of the two litters comprising each of the two treatment groups sampled at postnatal day 20 (the first sampling time point for DR and FF piglets) and postnatal day 154 (final time point). PCA plots generated using a dataset of 346 ASVs identified as having ≥0.1% relative abundance in at least 33% of fecal samples collected from pigs in a given treatment group at a given time point, revealed no significant effects of litter (P>0.05, PERMANOVA, Fig. S6; Dataset S7). We then used EMMER to select information-rich ASVs in the two litters comprising each treatment group at each of the two-time points sampled. PCA plots of the relative abundances of the ASVs present in the resulting eight lists of information-rich ASVs (Dataset S7**)** showed no significant effects of litter membership (P>0.05; PERMANOVA, Fig. S6). Finally, indicator species analysis (indicspecies, v1.7.9; ref. *3*) failed to identify statistically significant indicator species that could differentiate the microbiota of members of two litters of animals comprising a given treatment group at a given time point. Based on these results, we concluded that litter membership did not significantly influence microbiota configuration in our study.

EMMER versus Random Forests analysis of longitudinally sampled FF and DR pig microbiota

Applying EMMER to the V4-16S rDNA amplicon sequencing dataset (all time points) of 564 ASVs identified in the fecal microbiota at ≥0.1% relative abundance and in at least 33% of samples collected from animals in a given treatment group at a given time point, yielded 33 information-rich ASVs with high reproducibility in the FF dataset and 47 in DR dataset (median reproducibility; 92.3% in the FF dataset, 94.1% in the DR dataset; Dataset S2). The FF and DR datasets shared 10 information-rich ASVs. Together, these 70 information-rich ASVs represent a collection of ASVs that captures key characteristics of the two treatment groups at each sampling time point. Plotting the centroids of FF and DR microbiota at each time point on PCA revealed a temporal pattern of separation between the FF and DR groups that is first evident at postnatal day 49 (Fig. S2*G*). PERMANOVA indicated that DR and FF microbial community structures became significantly different after postnatal day 77 and that this difference was maintained through to the end of the experiment (Fig. S2*H*).

We then used Random Forests (RF) to compare microbiota development in the serially sampled FF and DR animals. To avoid potential effects from rarefication, we normalized our V4-16S rDNA amplicon dataset with DESeq2 (*4*) before applying RF (*5*). We trained the RF-derived model by using DESeq2 normalized 16S rDNA amplicon sequencing data from four of the 13 pigs in the FF group (R package 'randomForest', ntree=10,000; ref. *6*), thereby obtaining a set of 'age-discriminatory' ASVs. Our RF-derived model was reciprocally cross-validated by calculating the Pearson correlation r between predicted microbiota age (defined by the abundances of age-discriminatory strains at a given postnatal developmental time point) and the chronological age of the pig. We then applied the RF-derived model to DESeq2 normalized 16S rDNA amplicon sequencing data from all the pigs in the DR group to describe the stage of their microbiota development compared to chronologically-aged matched FF pigs. We then repeated the previous steps by iterating through all possible combinations of 4 animals for the training set (randomly choose 4 from 13 FF animals, yielding 715 training datasets to produce the sparse RF-derived models). The results revealed that in FF animals, microbiota age was significantly positively correlated with chronological age throughout the first 6 months of postnatal life [Pearson's $r = 0.92 \pm 0.02$ (mean ± SD)]. However, in DR animals, microbiota age only exhibited a statistically significant positive correlation with chronological age prior to postnatal day 77 [Pearson's $r = 0.89 \pm 0.02$ (mean ± SD) compared to 0.07 ± 0.07 after postnatal day 77; see Fig. S7]. In summary, both EMMER and RF demonstrated that long-term diet restriction results in a perturbed microbiota but with RF we needed to make arbitrary decisions about how to subset the dataset. This was not necessary with EMMER, where we could assess the statistical significance of differences between FF and DR microbial community structures with PERMANOVA at each time point sampled (Fig. S2*H*).

SUPPLEMENTARY REFERENCES

1. G. Strang. Introduction to Linear Algebra, Fifth Edition (Wellesley-Cambridge Press, Wellesley, MA, 2016).

2. N. W. Griffin, P. P. Ahern, J. Cheng, A. C. Heath, O. Ilkayeva et al., (2017). Prior dietary practices and connections to a human gut microbial metacommunity alter responses to diet interventions. *Cell Host & Microbe* **21**, 84-96.

3. M. D. Cáceres, P. Legendre, and M. Moretti (2010). Improving indicator species analysis by combining groups of sites. *Oikos* **42**, 679-698.

4. M. I. Love, W. H, S. Anders. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550.

5. P. J. McMudie and S. Holmes (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10,** e1003531.

6. A. Liaw, M. Wiener (2002). Classification and regression by randomForest. *R News* **2/3**, 18-22.

SUPPLEMENTARY FIGURES

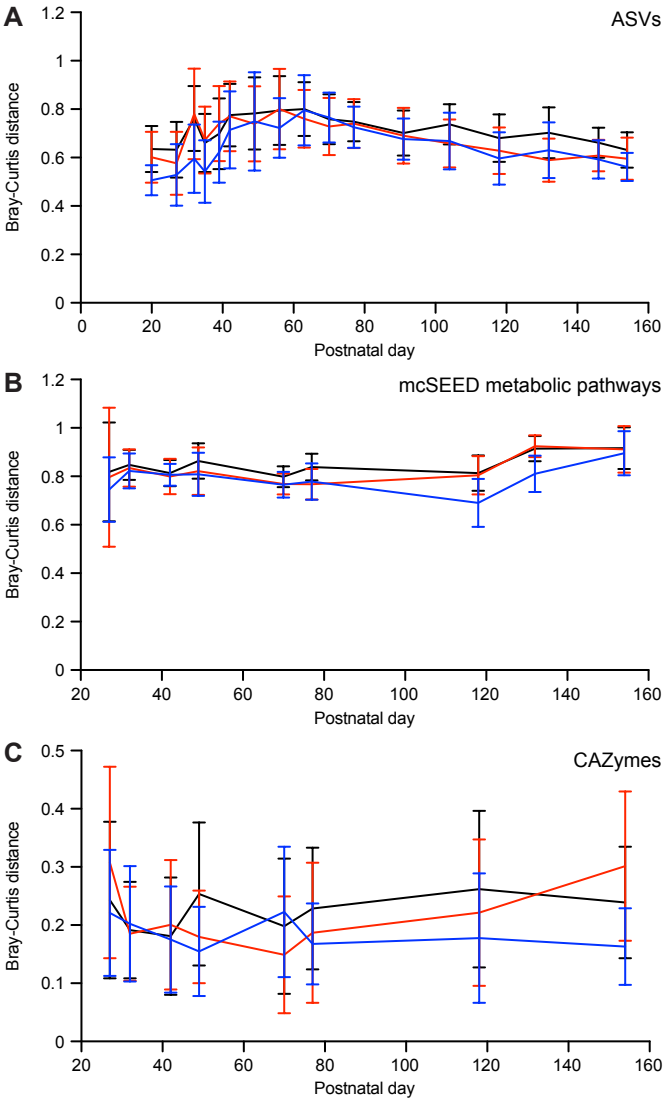

Fig. S1 – **β-Diversity analysis of pig fecal microbial communities. (A)** 16S rDNA-based analysis of ASV content **(B-C)** Microbiome diversity measurements based on the representation of genes in 104 mcSEED metabolic pathways (panel B) and genes encoding CAZymes (panel C). Bray-Curtis distances between each pair of samples within the DR group and each pair within the FF group are colored red and blue, respectively. Bray-Curtis distances between any possible combination of one DR sample and one FF sample at a given time point are shown in black. Error bars represent the standard deviation.

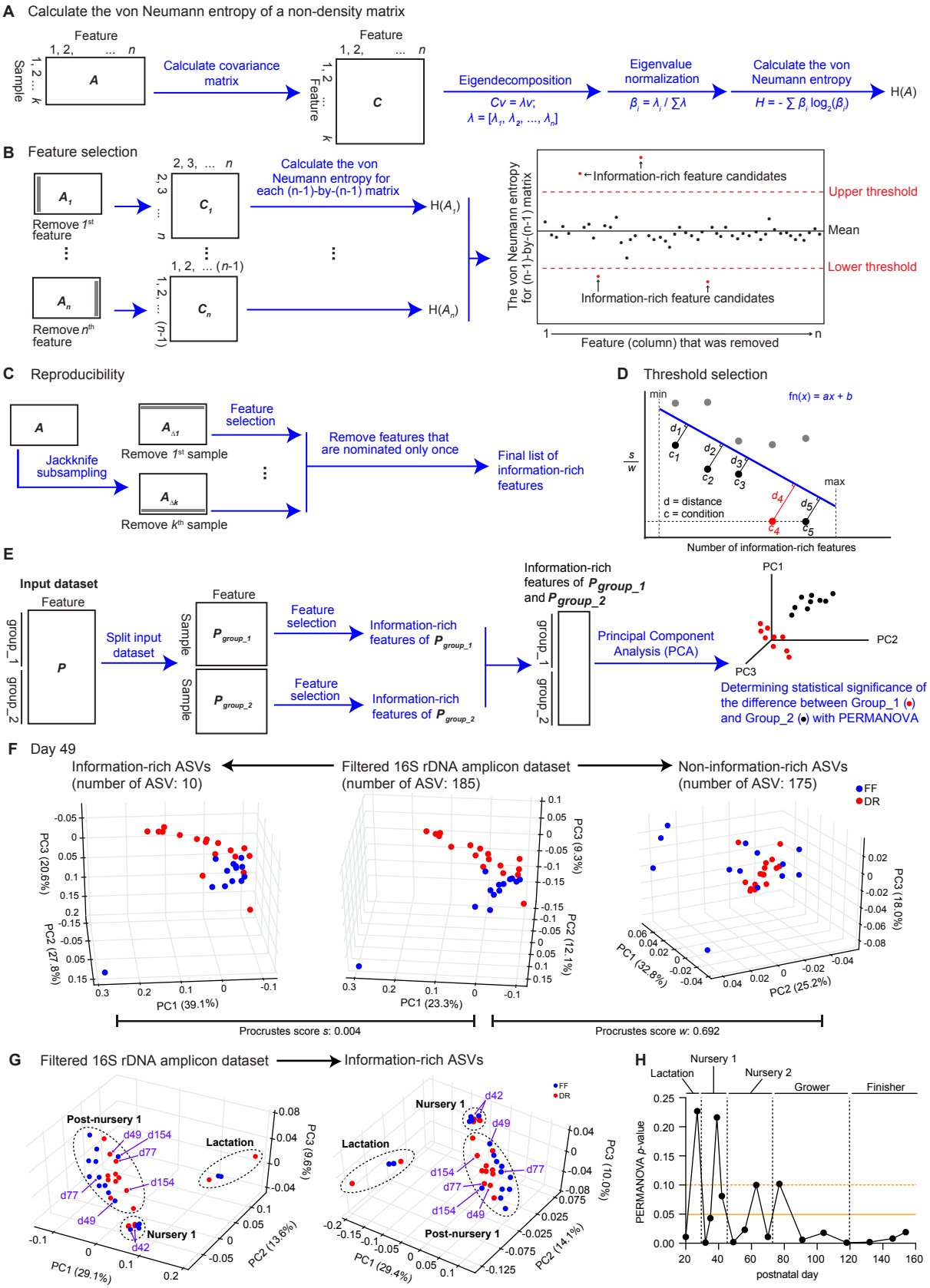

Fig. S2 – Identification of information-rich features in the gut microbiome/microbiota using EMMER. (A-D) Summary of the workflow for calculating the vNE of a non-density matrix, performing feature selection, evaluating the reproducibility of information-rich feature calling, and

optimizing the threshold for identifying information-rich feature. A table (matrix *A*) is constructed where columns are the abundances of a given feature type (CAZyme or ASV) and rows are animals in the same diet treatment group at a given time point. A covariance matrix (*C*) is then calculated. Eigendecomposition of matrix *C* yields eigenvalues that are normalized and used to calculate the vNE (panel A). Information-rich features are identified by systematically removing CAZymes or ASVs across the matrix (panel B) and verified by jackknife subsampling (panel C). The thresholds for selecting information-rich features are optimized by the method described in panel D. **(E)** Iterative strategy for identifying information-rich features in each treatment group. **(F)** In the example shown, dividing the postnatal day 49 microbiota dataset (n=185 ASVs) into two subsets based on diet treatment discloses that the optimal upper and lower thresholds are 1.5 units, yielding a total of 10 information-rich ASVs (7 from DR, 5 from DR and 2 that are shared; Dataset S2). Using these thresholds, the Procrustes score *s* (0.004) is considerably lower than the Procrustes score *w* (0.692) (Dataset S2)**. (G,H)** Centroids of the fecal microbiota configurations of members of each treatment group at each time point before and after EMMER feature selection (Dataset S2). A clear change occurred between postnatal day 42, the last time point sampled during consumption of the nursery 1 diet, and postnatal day 49, the first time point sampled for the nursery 2 diet (panel G). Projecting DR and FF animals onto PCA space with the representation of information-rich ASVs and applying PERMANOVA reveals that the differences between groups were sustained beyond postnatal day 77 (panel H), the time point when a 45% weight difference had been achieved.

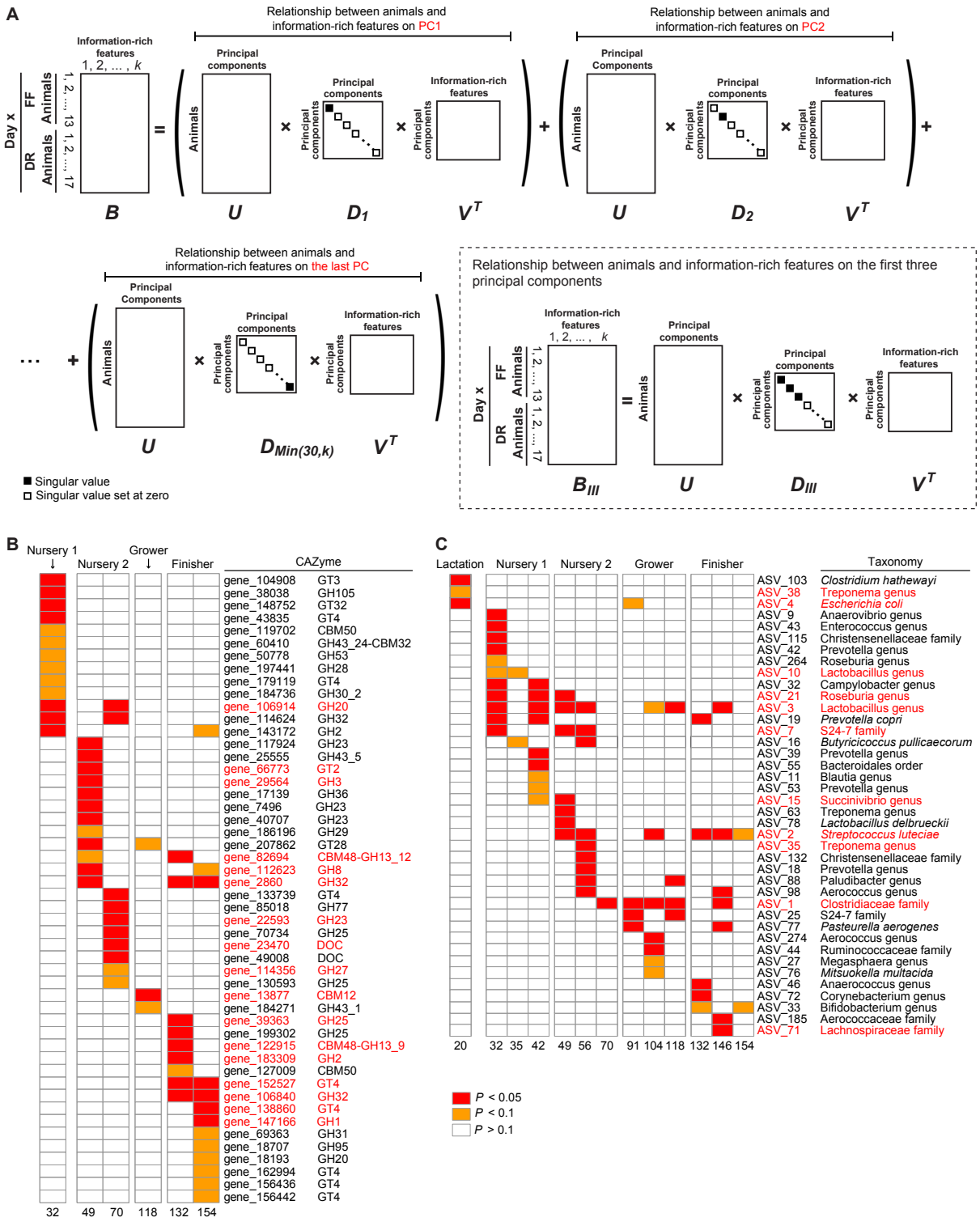

Fig. S3 – Identification of information-rich features that differentiate the gut communities of FF and DR pigs in PCA space. (A) The relationship between animals and information-rich features on a specific principal component (PC) can be isolated from the input matrix by singular value decomposition (SVD). **(B,C)** CAZymes (panel B) and ASVs (panel C) whose representation differentiate DR and FF pig microbiomes/microbiota on PCA plots at different time points. *P* values were calculated using linear regression.

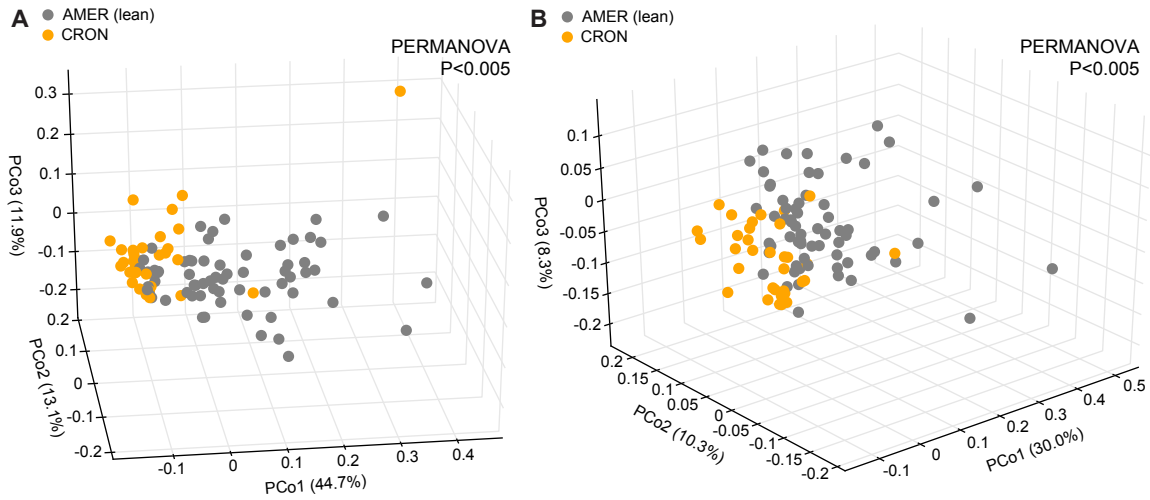

Fig. S4 – Testing EMMER using a published 16S rDNA amplicon dataset that contains fecal microbiota from groups of adults practicing chronic caloric restriction with adequate nutrition (CRON) or consuming a typical unrestricted USA diet (AMER). Principal Coordinates Analysis (PCoA) of unweighted UniFrac distances calculated based on the abundances of 242 DP-associated taxa (panel A), and 72 information-rich taxa (panel B).

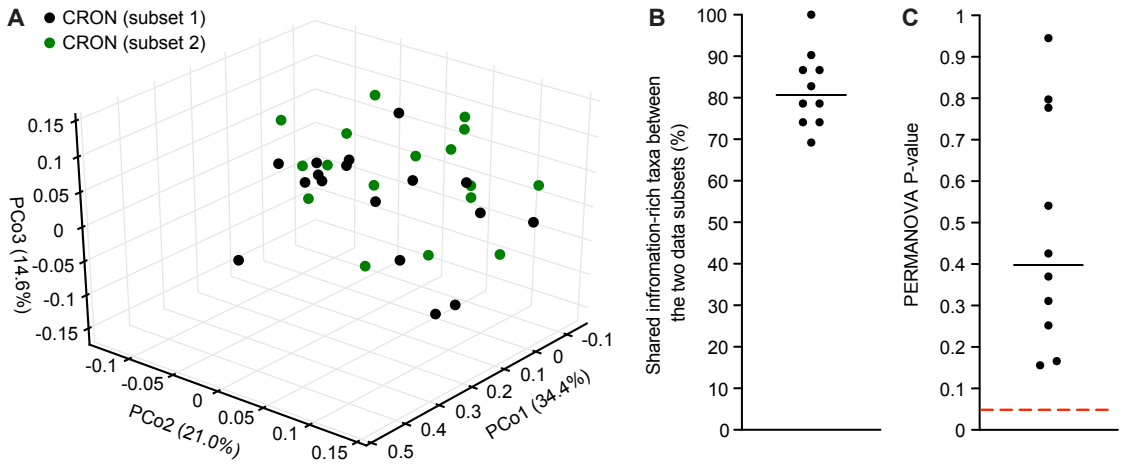

Fig. S5 – EMMER can be applied to datasets with similar gut microbial community configurations. (A) PCoA of unweighted UniFrac distances between fecal microbiota samples calculated based on the information-rich taxa identified from two CRON 16S rDNA amplicon sequencing data subsets (P>0.05; PERMANOVA). **(B,C)** Shared information-rich bacterial taxa (panel B) and PERMANOVA P-values (panel C) between subsets 1 and 2, calculated from each of the 10 iterations of randomly split CRON 16S rDNA amplicon data. Solid horizontal lines in panel B and C denote the mean, while the red dashed line in panel C indicates *P*=0.05. Each dot represents the result from one iteration.

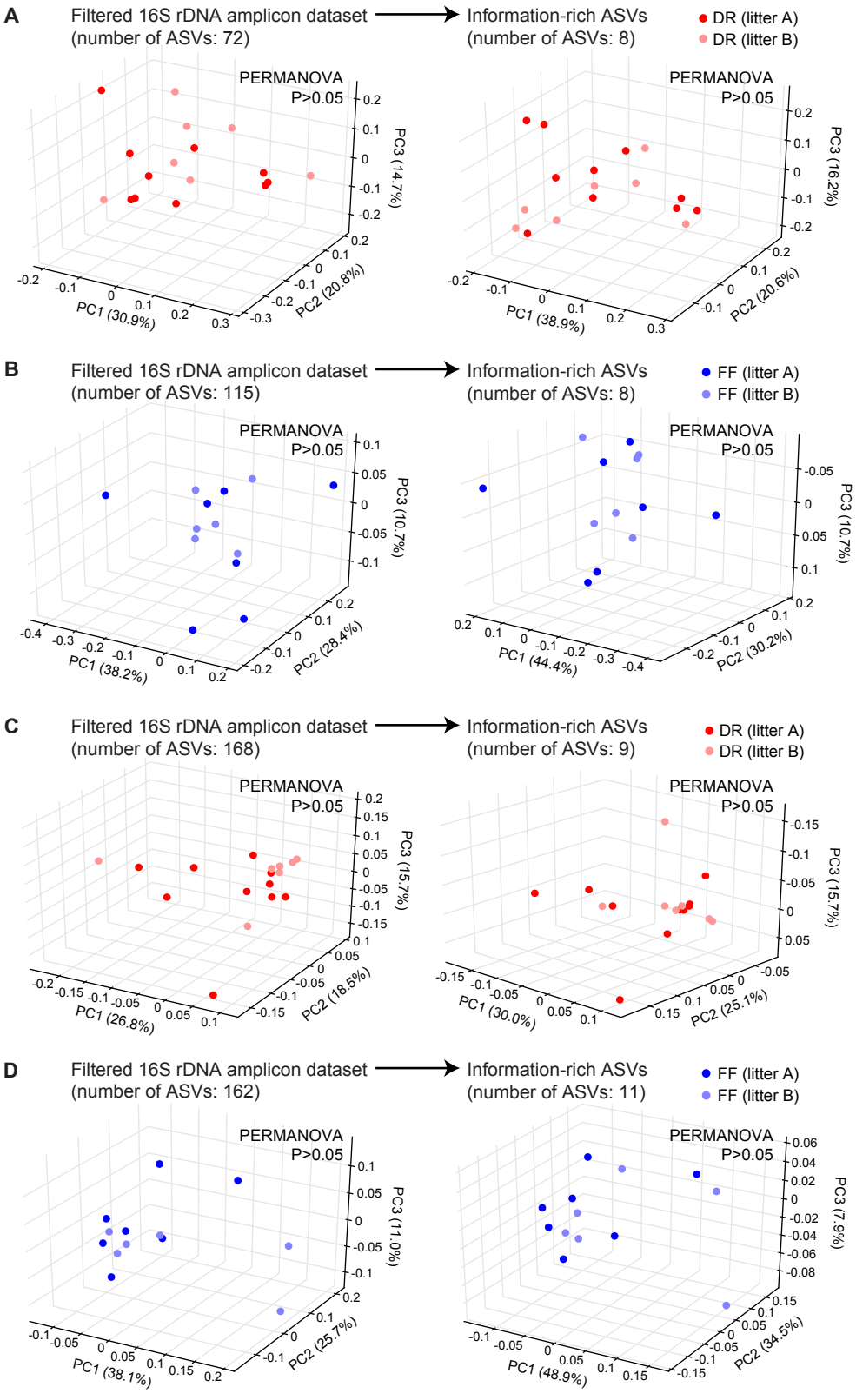

Fig. S6 – Assessing whether litter membership affects pig microbiota configuration. (A,B) PCA plots of the relative abundances of ASVs inputed into EMMER, and plots of the relative abundances of information-rich taxa selected by EMMER in the fecal microbiota of pigs belonging to the two litters comprising the DR or FF treatment groups. Results for fecal samples collected on postnatal day

20 are shown. **(C,D)** PCA plots generated using the relative abundances of ASVs in fecal samples collected on postnatal day 154.

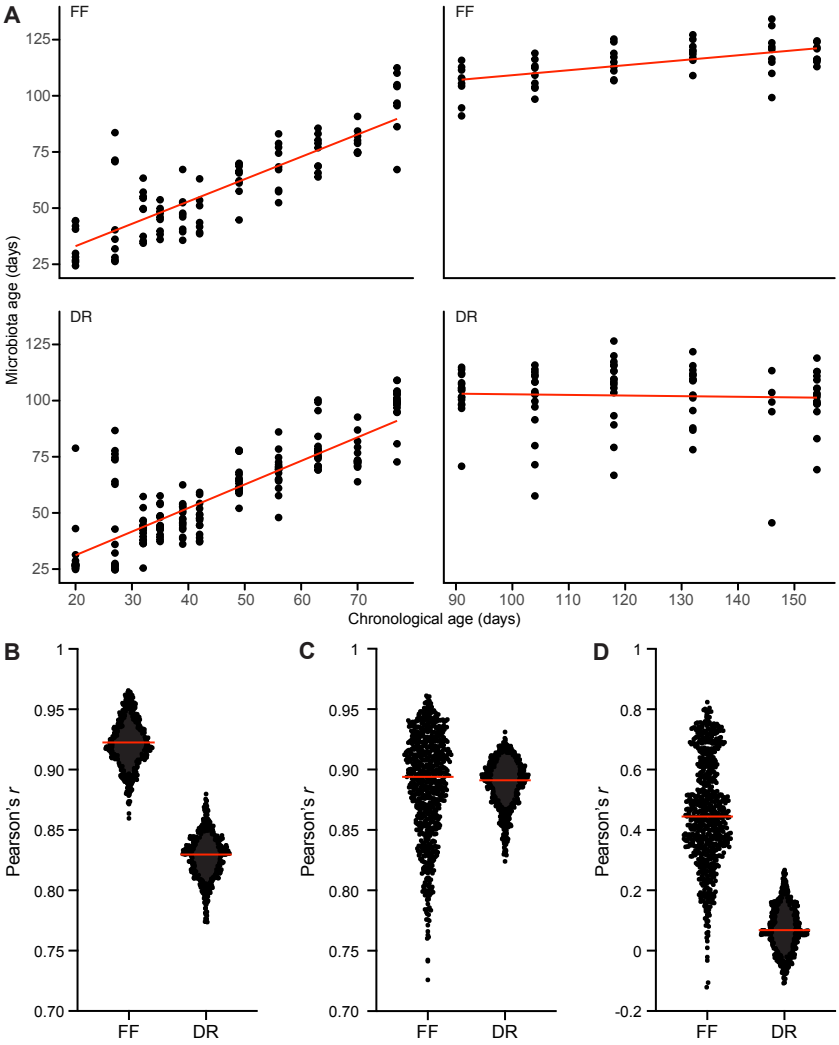

Fig. S7 – Using Random Forests (RF) to characterize microbiota development in DR and FF pigs. (A) Applying a sparse RF-derived model, trained from four FF animals, to 16S rDNA amplicon sequencing datasets generated from the remaining nine FF animals (top panel) and all 17 DR animals (lower panel). After postnatal day 77, microbiota age is no longer significantly positively correlated with chronological age in DR pigs. **(B-D)** 715 iterations through all possible combinations of 'choose 4 pigs from all 13 FF animals' for the training dataset. Pearson's *r* was calculated between microbiota age and chronological age for all time points (panel B), time points before postnatal day 77 (panel C) and time points after postnatal day 77 (panel D); each dot represents the result obtained from one iteration. Red horizontal lines in panels B-D indicate mean values.

SUPPLEMENTARY DATASETS

Dataset S1 – Characterization and consumption of diets. (A) Composition. **(B)** Nutrient analysis. **(C)** Food consumption.

Dataset S2 – Growth, metabolic and microbiome characteristics of DR and FF pigs. (A) Growth phenotypes. **(B)** Microbiome characteristics. **(C)** Metabolic characteristics.

Dataset S3 – Pig fecal microbiomes used in gnotobiotic mouse experiments.

Dataset S4 – Characterizing gnotobiotic mice in the *ad libitum* feeding experiment. (A) CAZyme gene content in cecal and fecal microbiomes. **(B)** Levels of metabolites.

Dataset S5 – Characterizing gnotobiotic mice in the controlled feeding experiment. (A) Weights. **(B)** CAZyme gene content in cecal and fecal microbiomes. **(C)** Levels of metabolites.

Dataset S6 – **Applying EMMER to a published dataset. (A)** Reproducibility of information-rich taxa calling. **(B)** PCoA coordinates used to generate **Fig. S3A**. **(C)** PCoA coordinates used to generate **Fig. S3B**. **(D)** PCoA coordinates used to generate **Fig. S4A**.

Dataset S7 – **Assessment of litter membership on pig microbiota configuration. (A)** Reproducibility of information-rich ASV calling. **(B)** PCA coordinates used to generate **Fig. S5.**