nature | methods

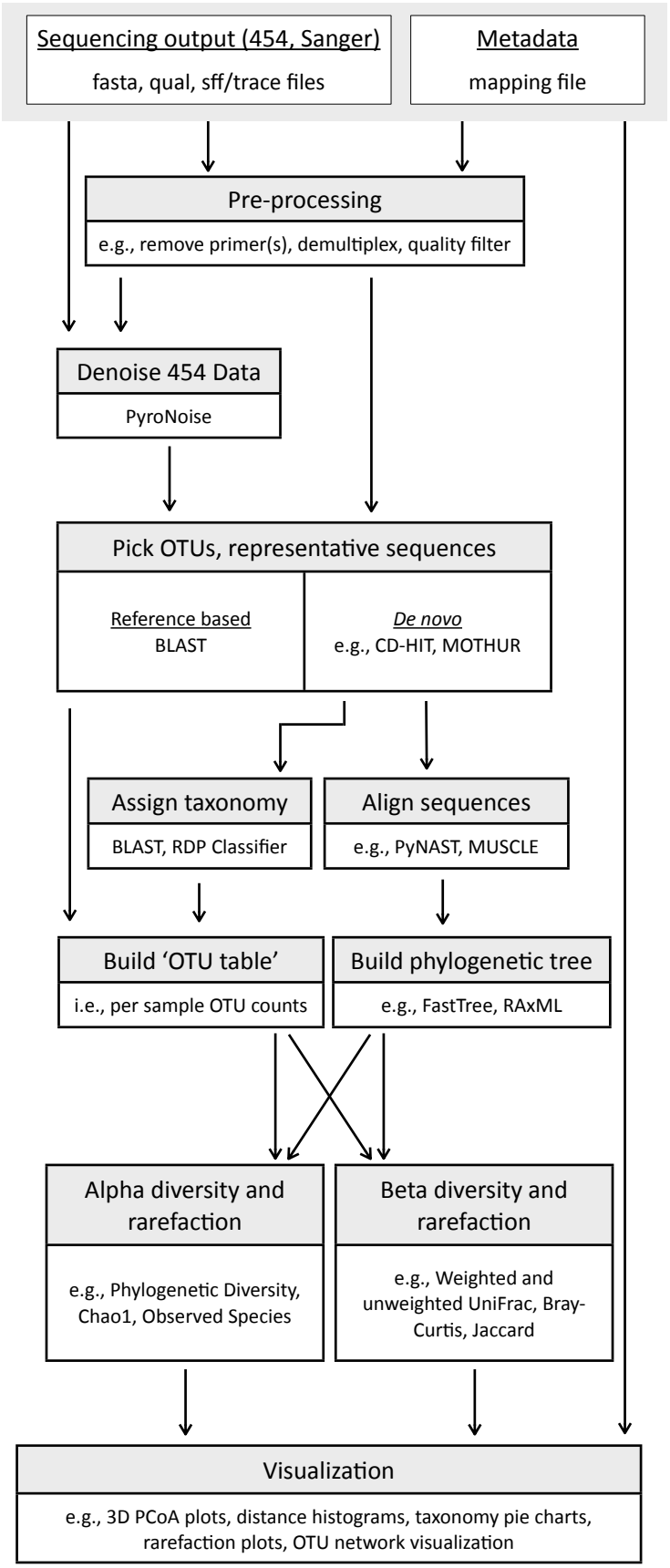
QIIME allows analysis of high-throughput community sequencing data

J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld & Rob Knight

Supplementary figures and text:

Supplementary Figure 1	Overview of the analysis pipeline.
Supplementary Table 1	Details of conventionally raised and conventionalized mouse samples.
Supplementary Discussion	Expanded discussion of QIIME analyses presented in the main text; Sequencing of 16S rRNA gene amplicons; QIIME analysis notes; Expanded Figure 1 legend; Links to raw data and processed output from the runs with and without denoising.

Supplementary Figure 1. Overview of the analysis pipeline.



Supplementary Figure 1. Overview of the analysis pipeline. The QIIME workflow begins with raw sequencing data plus metadata describing the samples, and can provide tabular output including many alpha and beta diversity measures in addition to publication-quality graphics. This figure illustrates only a subset of QIIME's core capabilities. Notable features include flexibility in terms of algorithmic choices at different steps, ease of deployment in cluster and other multi-processor environments, and modularity. QIIME is not limited to analyses based on the 16S rRNA gene: it can be used for any collection of sequences, although taxonomy assignment can only be used in cases where a reference database of sequences with assigned taxonomies is available. The components of the displayed workflow are as follows: first, data from the sequencing instrument and the metadata supplied by the user are combined to de-multiplex the barcoded reads from the various samples, and to perform quality filtering. Second, for 454 data, denoising can be performed using PyroNoise. Third, the sequences are grouped onto OTUs (Operational Taxonomic Units) at a user-defined level of sequence similarity (e.g. 97% to approximate species-level phylotypes). This step can be performed either using a reference database of OTU representatives (e.g. with BLAST), or purely based on sequence similarity (e.g. using uclust, cd-hit, or MOTHUR). Fourth, once the OTUs are picked and the representative sequences chosen, taxonomy is assigned, the sequences are aligned, and phylogenetic trees are built (again, multiple choices are possible at each of these steps: for example, the trees can be built *de novo* from the reference sequences using software such as FastTree or RAxML, or they can be built by mapping reads to their relatives in a predefined reference tree using BLAST. At this stage, a table showing the counts of each OTU in each sample is also produced. Sixth, the OTU tables are used to perform alpha and beta diversity calculations (alpha diversity refers to the diversity within each sample, and beta diversity refers to patterns of similarity and difference among samples). Finally, the alpha and beta diversity measurements are combined with metadata about each sample, potentially including new metadata supplied by the user at this step, to produce visualizations that allow the information to be readily interpreted.

Supplementary Table 1: Details of conventionally-raised (CONV-R) and conventionalized (CONV-D) C57BL/6J male mice samples.

Sample	Barcode Sequence	Diet	Colonization state	Sequences
MD2	ATGAGACTCCAC	Western	CONV-R	1067
MD4	ATGTGCACGACT	Low-carb	CONV-R	853
MD6	CACATCTAACAC	Western	CONV-R	1220
MD7	CAGACTCGCAGA	Western	CONV-R	896
MD8	CAGTGATCCTAG	Low-carb	CONV-R	930
MD9	CATGAGTGCTAC	Low-carb	CONV-R	772
MD18	ATACGTCTTCGA	Low-fat	CONV-R	901
MD19	ATCGATCTGTGG	Low-fat	CONV-R	850
MD20	ATGATCGAGAGA	Western	CONV-R	940
MD21	ATGTGTCGACTT	Low-carb	CONV-R	720
MD22	CACATTGTGAGC	Western	CONV-R	0
MD23	CAGAGGAGCTCT	Low-carb	CONV-R	1131
MD24	CAGTGCATATGC	Low-fat	CONV-R	776
MD25	CATGCAGACTGT	Low-fat	CONV-R	1025
Myd1	CAGTCACTAACG	LF/PP	CONV-R	1199
Myd2	CATCGTATCAAC	LF/PP	CONV-R	1415
Myd3	ATACACGTGGCG	LF/PP	CONV-R	1301
Myd4	ATCCGATCACAG	Western	CONV-R	951
Myd5	ATGACTCATTCG	Western	CONV-R	937
Myd6	ATGTCACCGTGA	Western	CONV-R	1216
Rag1	CACAGTGGACGT	LF/PP	CONV-R	20
Rag2	CAGACATTGCGT	LF/PP	CONV-R	1119
Rag3	CAGTCGAAGCTG	LF/PP	CONV-R	1083
Rag4	CATCTGTAGCGA	Western	CONV-R	1061
Rag5	ATACAGAGCTCC	Western	CONV-R	1209
Rag6	ATCCTCAGTAGT	Western	CONV-R	1051
WD2	CACACGTGAGCA	LF/PP	CONV-D	453
WD3	CACTGGTATATC	LF/PP	CONV-D	1997
WD4	CAGTACGATCTT	LF/PP	CONV-D	501
WD5	CATCATGAGGCT	LF/PP	CONV-D	767
WD6	ATAATCTCGTCG	Western	CONV-D	1378
WD7	ATCAGGCGTGTG	Western	CONV-D	970
WD8	ATGACCATCGTG	Western	CONV-D	458
WD9	ATGTACGGCGAC	Western	CONV-D	1273
WD10	CACAGCTCGAAT	Western	CONV-D	684
WD11	CACTGTAGGACG	LF/PP	CONV-D	113

Supplementary Discussion

Expanded discussion of QIIME Analyses presented in the main text:

Here we illustrate how the QIIME workflow can be applied to a meta-analysis of three independent studies of distal gut bacterial communities. The first study[1] evaluated gut communities from adult human monozygotic and dizygotic twins and their mothers. The second study[2] evaluated gut communities from germ-free mice and germ-free conventionalized mice (GF and CONV-D mice), where CONV-D mice are germ-free mice colonized with a mouse gut microbiota from conventionally-raised mouse donors. The third study[3] evaluated gut communities from a time-series of adult gnotobiotic mice after they received a human fecal microbiota that had been transplanted by oral gavage; the human donors; and conventionally-raised control mice (CONV-R, or animals exposed to a gut microbiota from their mothers and environment starting at birth) and CONV-D control mice (**Supplementary Table 1**). This analysis combines ten full 454 FLX runs and one partial run, totalling 3.8 million bacterial 16S rRNA sequences from previously published studies: it also includes reads from different regions of the 16S rRNA gene (variable region 2 (V2) versus variable region 6 (V6)).[1] A step-by-step guide to the QIIME analysis can be found in the 'QIIME commands' section of this document. A smaller analysis, suitable for running on a laptop, can be found with the QIIME tutorial at http://qiime.sourceforge.net.

Several results are immediately apparent from the principal coordinates (PCoA) plots based on UniFrac distances (**Fig. 1a**), in which samples that cluster together have similar bacterial community membership. First, the distal gut (cecal) microbiota of CONV-R mice from one study cluster together with the cecal microbiota from CONV-R mice with the same genetic background and diet from another study, and from conventionalized (CONV-D) animals. Second, QIIME illustrates findings, described previously[3], from a time series study of fecal samples obtained from gnotobiotic mice 8 hours through 56 days after they received a human fecal microbiota transplanted with a single oral gavage. During the first week after transplantation, fecal

microbiota structure rapidly evolves towards that of the human fecal donor sample, whether the transplanted sample was fresh, or had been frozen for a year prior to transplantation. Fecal community configuration is sustained from Day 7 through Day 56. The transplanted gut community is highly sensitive to host diet, as judged from a comparison of mice on a standard low-fat and high plant polysaccharide (LF and PP) chow or those switched to a high fat/high sugar western diet. QIIME allows this data to be placed in the broader context of data obtained from the fecal microbiota of members of >50 families of adult twins and their mothers, described in a previous study[1], and data from CONV-R mice generated by another research group. Third, analysis of histograms representing all pairwise UniFrac measurements of phylogenetic distances between samples (**Fig. 1b**), indicates that distances between the fecal microbiota of humanized gnotobiotic mice and the fecal microbiota of human twins decrease as a function of the number of days following transplant of the human donor's microbiota. Only a subset of the humanized mice time-points are included in **Fig. 1b** to facilitate visualization of this point. Taxon-based measures, which can also be computed by QIIME, give qualitatively similar results (data not shown). The taxonomy pie charts shown (**Fig. 1b**) indicate that the humanized gnotobiotic mice sampled at 8h on the first day (Day 0) are primarily dominated by Erysipelotrichi (a class within the Firmicutes), while adult human twins are primarily dominated by Clostridia (another Firmicutes class) and the Bacteroidetes**.** As the transplanted human microbiota stabilizes in the mouse gut, we see a shift toward the taxonomic representation encountered in human communities. Thus, QIIME documents a shift towards a human-like community using several types of analyses: the location of each sample on the PCoA plot (**Fig. 1a**), distance histograms (**Fig. 1b**), inspection of the taxonomy pie charts (**Fig. 1b**), as well as convergence in overall diversity (**Fig. 1c**).

The vast majority of the data described above were generated with reads from the V2 region of the 16S rRNA genes. The human twin fecal microbiota data show that the difference between V2 and V6 reads obtained from the same samples are comparable to the distances between

mouse and human gut microbiota, suggesting that the effects of primers used for PCR of 16S rRNA genes can be large (**Fig. 1a**).

One emerging concern in the analysis of pyrosequencing and other high-throughput data is the effect of sequencing noise[4], and QIIME therefore supports denoising of pyrosequencing data. Phylogenetic Diversity (PD) rarefaction curves for raw and denoised humanized mice data were used to evaluate the affect of denoising (**Fig. 1c**). Denoising was performed using a custom implementation of the PyroNoise[4] algorithm (manuscript in preparation). Transplantation of a human gut microbiota into germ-free mice results in a rapid increase in PD as human gut microbes colonize the mouse intestine. At Day 7, the communities have largely stabilized. In both the raw and the denoised data, the relative PD follow the same trend but denoising reduces this measure of alpha diversity by a factor of two (for rarefaction curves see **Fig. 1c**). Alpha diversity measures are affected more severely than are the beta diversity measures by noise, and phylogenetic beta diversity measures such as UniFrac are especially robust. The improved performance of phylogenetic methods over taxon-based methods in general is expected because the new OTUs introduced by noise are not so different from existing sequences that they cannot be related to existing parts of the tree. This is related to, but separate from, the way that phylogenetic beta diversity measures (unweighted unifrac in this case) eliminate the “spike” artifacts at 90 degree angles (**Fig. 1a**): these stem from the high levels of dissimilarity between samples at the species level, and obscure the clustering patterns that the phylogenetic measures reveal.

The meta-analysis described above took ~12h on a 100 processor Linux cluster without denoising. The majority of time (approximately 11 of the 12h) on this data was spent picking OTUs, which was done via BLAST to facilitate integration of 454 pyrosequencing data from non-overlapping regions of the 16S rRNA gene (V2 and V6). Denoising required approximately ~120h on a 100 processor Linux cluster, and downstream analyses of the resulting denoised data required ~1h on the same Linux cluster. Most of the steps related to visualization are rapid:

for example, principal coordinates reduction of the UniFrac matrix and generation of the 3-dimensional plots and histograms each take about a minute on a laptop. Complete analyses of smaller datasets, such as a partial 454 run, can be done on a laptop in a few hours.

Availability:

QIIME is open source, and available from sourceforge at http://qiime.sourceforge.net. An extensive tutorial, and the raw input data and the processed data from the analyses presented below, are available via http://qiime.sourceforge.net.

Sequencing methods:

Sequencing of 16S rRNA gene amplicons – Mouse cecal samples from a set of diet switch experiments were stored at -80˚C before processing[5]. DNA was extracted by bead beating followed by phenol-chloroform extraction as described previously[1]. The V2 region was targeted for amplification by PCR (with primers 8F-338R) and multiplex FLX pyrosequencing. See **Supplementary Table 1** for a list of the sequenced samples.

Sequencing of amplicons for previously published studies is described in the corresponding publications.

Overview of technology used:

QIIME is implemented using Python 2.6 and the PyCogent toolkit[6]. It relies on the Python libraries numpy (http://numpy.scipy.org), matplotlib (http://matplotlib.sourceforge.net) and (optionally) cython (http://www.cython.org). It wraps a number of third-party applications including BLAST[7], MOTHUR[8], DOTUR[9], and cd-hit[10] for OTU picking, MUSCLE[11], Clustal[12], MAFFT[13] and DIALIGN[14] for alignment, the RDP classifier[15] for taxonomy assignment (BLAST can also be used for this task), PyroNoise[4] for denoising, and Cytoscape[16] and KiNG (http://kinemage.biochem.duke.edu) for visualization (we also recommend FigTree, http://tree.bio.ed.ac.uk/software/figtree/, TopiaryExplorer

(http://sourceforge.net/projects/topiarytool/), and PyCogent for visualizing phylogenetic trees and cluster diagrams).

QIIME commands:

This section presents the QIIME commands that were used in the data analyses presented herein. QIIME commands are presented in `monospace font`.

Qiime paper analysis with V2/V6 data combined, raw (i.e., not denoised) data

Set up environment

```
$q=/home/qiime/Qiime/qiime $working_dir=/home/qiime/Hmice_Raw_w_V6/
$refdb=/home/qiime/greengenes_unaligned.fasta-OTUs_at_0.01.fasta
$tree=/home/qiime/greengenes_lanemasked_filtered.ntree
```

Pick OTUS

```
python $q/parallel/pick_otus_blast.py -i $working_dir/seqs.fna -o
$working_dir/blast_picked_otus/ -r $refdb -O 100 -e 1e-20
```

Pick representative set

```
python $q/pick_rep_set.py -i
$working_dir/blast_picked_otus/seqs_otus.txt -f $working_dir/seqs.fna
-o $working_dir/repr_set.fasta
```

Assign taxonomy

```
python $q/parallel/assign_taxonomy_rdp.py -i
$HOME/Hmice_Raw_w_V6/repr_set.fasta -o $HOME/Hmice_Raw_w_V6
```

Build OTU table

```
python $q/make_otu_table.py -i
$working_dir/blast_picked_otus/seqs_otus.txt -t
$working_dir/repr_set_tax_assignments.txt -o
$working_dir/otu_table.txt
```

Perform Alpha Diversity on OTU Table

```
python $q/alpha_diversity.py -t $tree -m osd,PD_whole_tree -i
$working_dir/otu_table.txt -o $working_dir/alpha_osd_PD.txt
```

Perform Beta Diversity on OTU Table

```
python $q/beta_diversity.py -t $tree -m dist_unweighted_unifrac -i
$working_dir/otu_table.txt -o $working_dir/beta_unweighted_unifrac.txt

python $q/beta_diversity.py -t $tree -m dist_weighted_unifrac -i
$working_dir/otu_table.txt -o $working_dir/beta_weighted_unifrac.txt

python $q/beta_diversity.py -t $tree -m dist_euclidean -i
$working_dir/otu_table.txt -o $working_dir/beta_dist_euclidean.txt
```

Generate a Coords file for Beta Diversity

```
python $q/principal_coordinates.py -i
$working_dir/beta_unweighted_unifrac.txt -o
$working_dir/beta_unweighted_unifrac_coords.txt
```

```
python $q/principal_coordinates.py -i
$working_dir/beta_weighted_unifrac.txt -o
$working_dir/beta_weighted_unifrac_coords.txt

python $q/principal_coordinates.py -i
$working_dir/beta_dist_euclidean.txt -o
$working_dir/beta_dist_euclidean_coords.txt
```

Generate 3D Plots using the beta-diversity coords file and mapping file

```
python $q/make_3d_plots.py -i
$working_dir/beta_unweighted_unifrac_coords.txt -o
$working_dir/3d_plots/ -p $working_dir/Qiime_paper_prefs_filter.txt -m
$working_dir/Mice_Hmice_Twins_Mapping_Plots.txt

python $q/make_3d_plots.py -i
$working_dir/beta_weighted_unifrac_coords.txt -o
$working_dir/3d_plots/ -p $working_dir/Qiime_paper_prefs_filter.txt -m
$working_dir/Mice_Hmice_Twins_Mapping_Plots.txt

python $q/make_3d_plots.py -i
$working_dir/beta_dist_euclidean_coords.txt -o $working_dir/3d_plots/
-p $working_dir/Qiime_paper_prefs_filter.txt -m
$working_dir/Mice_Hmice_Twins_Mapping_Plots.txt
```

Alpha rarefaction

```
python $q/parallel/rarefaction.py -i $working_dir/otu_table.txt -o
$working_dir/alpha_rare/ -m 100 -x 2500 -s 250 -n 10 -N
$q/rarefaction.py

python $q/parallel/alpha_diversity.py -t $tree -m chao1,PD_whole_tree
-i $working_dir/alpha_rare -o $working_dir/rare_chao1_PD -N
$q/alpha_diversity.py

python $q/collate_alpha.py -i $working_dir/rare_chao1_PD/ -o
$working_dir/rare_collated
```

Create Rarefaction Plots

```
python $q/make_rarefaction_plots.py -m
../Qiime_paper_Full_mapping_All.txt -r PD_whole_tree.txt -o ./
```

Qiime paper analysis with V2/V6 data combined, denoised data

Note that the denoising steps use currently unpublished software, so cannot be run with QIIME alone (manuscript is in preparation). Smaller datasets can be handled with QIIME together with the published version of the PyroNoise software[4].
The denoiser was run separately on each of the 10 full GSFLX runs, and for each run we started with these input files in a separate folder:
454Reads.sff.txt
454Reads.fasta
454Reads.qual (from sffinfo)
Barcode_mapping.txt

Demultiplex and quality filtering, creates output seqs.fna, increase the number after -s by 1000000 to avoid collisions from the same sample_id in different runs (for V6 add: `-l 90 -L 110 -b 5 -p "CNACGCGAAGAACCTTANC,CAACGCGAAAAACCTTACC,CAACGCGCAGAACCTTACC,ATACGCGARGAACCTTACC,CTAACCGANGAACCTYACC").`

```
python split_libraries.py -m Barcode_mapping.txt -i 454Reads.fasta -q
454Reads.qual -r -s 1000000
```

Cut barcodes and primers from flowgrams, prefix dereplication

```
python denoise_preprocess.py -i 454Reads.sff.txt -f seqs.fna -o
Preprocessed/ -s -v
```

Run actual denoising with 40 cpus on the cluster. Produces output files:
centroids.fasta
unclustered.fasta
denoiser_mapping.txt

```
python denoiser.py -i 454Reads.sff.txt -p Preprocessed -o Denoised/ -c
-n 40 -b 3 -v
```

Combine centroids and singletons

```
cat Denoised/unclustered.fasta Denoised/centroids.fasta
>/Denoised/denoised.fasta
```

Finished denoising of run, repeat with next run
After all runs are denoised, combine results of all 10 runs.

```
cat  */Denoised/denoised.fasta > all_runs_denoised.fasta cat
*/Denoised/denoiser_mapping.txt > all_runs_denoiser_mapping.txt cat
*/seqs.fna > all_split_libraries_seqs.fna
```

Pick OTUs

```
python $q/parallel/pick_otus_blast.py -i
$HOME/qiime_paper_denoised/all_denoised.fasta -o
$HOME/qiime_paper_denoised/blast_picked_otus_1e-20 -r
$HOME/greengenes_filtered/greengenes_unaligned.fasta-
OTUs_at_0.01.fasta -O 100 -e 1e-20
```

Combine denoiser mapping and OTU mapping, convert flowgram id's to unique sample_ids taken from split_libraries.py output. Creates final OTU map (qiime_input_otu_map.txt) and replaces ids in denoised fasta file: qiime_input_seqs.fasta

```
python denoiser_to_qiime_linker.py all_split_libraries_seqs.fna
all_runs_denoiser_mapping.txt all_runs_denoised.fasta
blast_picked_otus/all_runs_denoised_otus.txt
```

Pick representative seqs (Note: the members of an otu are ordered, such that '-m first' on this otu map is basically identical to the usual qiime default '-m most_abundant') -- result is $HOME/qiime_paper_denoised/blast_picked_otus_1e-20/repr_set.fasta

```
python pick_rep_set.py -m first -i qiime_input_otu_map.txt -f
qiime_input_seqs.fasta
```

Assign taxonomy

```
python $q/parallel/assign_taxonomy_rdp.py -i
$HOME/qiime_paper_denoised/blast_picked_otus_1e-20/repr_set.fasta -o
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/ -O 100
```

Build OTU Table

```
python $q/make_otu_table.py -i
$HOME/qiime_paper_denoised/blast_picked_otus_1e-20/otus.txt -t
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/repr_set_tax_assignments.txt -o
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/otu_table.txt
```

Beta diversity

```
python $q/beta_diversity.py -i
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/otu_table.txt -m dist_unweighted_unifrac -o
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/beta_unweighted_unifrac.txt -t
$HOME/greengenes_filtered/greengenes_lanemasked_filtered.ntree

python $q/principal_coordinates.py -i
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/beta_unweighted_unifrac.txt -o
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/beta_unweighted_unifrac_coords.txt

python $q/beta_diversity.py -i
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/otu_table.txt -m dist_weighted_unifrac -o
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/beta_weighted_unifrac.txt -t
$HOME/greengenes_filtered/greengenes_lanemasked_filtered.ntree

python $q/principal_coordinates.py -i
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/beta_weighted_unifrac.txt -o
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/beta_weighted_unifrac_coords.txt

python $q/make_3d_plots.py -i
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/beta_unweighted_unifrac_coords.txt -m
Mice_Hmice_Twins_Mapping_Plots.txt -o
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/3d_unweighted/ -p
Qiime_paper_prefs_filter.txt

python $q/make_3d_plots.py -i
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/beta_weighted_unifrac_coords.txt -m
Mice_Hmice_Twins_Mapping_Plots.txt -o
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/3d_weighted/ -p Qiime_paper_prefs_filter.txt
```

Alpha Rarefaction/Diversity

```
python $q/parallel/rarefaction.py -i
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/otu_table.txt -o
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/rarefaction/ -m 100 -x 2500 -s 250 -n 10
python $q/parallel/alpha_diversity.py -i
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/rarefaction/ -o
$HOME/qiime_paper_denoised/blast_picked_otus_1e-
20/rdp_assigned_taxonomy/alpha_diversity/ -m
observed_species,chao1,PD_whole_tree -t
$HOME/greengenes_filtered/greengenes_lanemasked_filtered.ntree

python $q/collate_alpha.py -i ./alpha_diversity/ -o
./alpha_diversity_collated/
```

Expanded Figure 1 Legend. QIIME analyses of the distal gut microbiotas of conventionally-raised and conventionalized mice, gnotobiotic mice colonized with a human fecal gut microbiota (H-mice), and human adult mono- and dizygotic twins. The colors in the legend are used consistently throughout the panels, and separate samples by species and timepoint. (**a**) Principal coordinates analysis (PCoA) plot of mice, H-mice and twins using denoised sequence data and unweighted UniFrac (left), or Euclidean distances (right). The axes are PC1 (principal coordinate 1), PC2, and time in days. Human and mouse fecal samples were added at the end of the time series, to observe whether the stabilized H-mice communities were more similar to human twins or mice. The time series is shown as a rainbow color gradient from red (earliest time points) to cyan (latest time points); these contrast with the clear clusters of human (dark blue), mice (light and dark purple for the two studies), human fecal community 16S rRNA V6 reads (grey, rather than V2 reads). Note that the separation between communities correlated with the time series by UniFrac is lost when Euclidean distance is applied in the right panel: this result underscores the power of phylogenetic methods. (**b**) Histograms of unweighted UniFrac distances between the fecal microbiota of adult human twins, and Day 0 post-transplant H-mice on a low-fat/plant polysaccaride-rich diet (LF and PP) diet; Day 1 H-mice (LF and PP diet); Day 56 H-mice (LF and PP diet); human donor for the H-mice study; and human twins (i.e., within category distances), all using the denoised sequence data. The histogram plots the distribution of UniFrac distances for a given comparison: smaller distances indicate groups of samples that are on average more similar as shown by branch length in a phylogenetic tree. Only a subset of the H-mice time-points are included in Fig. 1b to facilitate non-interactive visualization: in the interactive results, more timepoints can be examined easily. The colors match the colors of the relevant samples in Fig. 1a, and pie charts (colors shown in pie chart legend) summarize the taxa in those sets of samples (not the differences between those samples and the starting point). The pie charts associated with each group reinforce the convergence of community configuration in gnotobiotic mouse recipients of the human donor's fecal microbiota towards the donor's community as a function of time after

colonization: note that the transplanted community begins with a configuration that is very different from that of both humans or conventionally-raised (CONV-R) mice. The series of histograms demonstrates that the successive timepoints diverge from the starting point after transplantation. (**c**) Alpha diversity rarefaction plots of Phylogenetic Diversity for the H-mice samples (Day 0 through Day 56 on LF and PP diet). Raw sequence data (left) is compared to denoised sequence data (right). Note that the raw sequence data has far more diversity (as represented by branch length on the tree) than the denoised sequence data, but the relative amounts of diversity and hence the conclusions are unchanged. Colors for the groups are the same as in the other panels: means and standard deviations of the rarefaction curves for individual samples in each group are shown. (**d**) Connectivity of H-mice time series data (Day 0 through Day 56 on LF and PP diet) using an OTU network, where lines connect the categories (or samples from the same time point and diet; top) and OTUs (bottom). This is a static representation of an interactive display in Cytoscape that allows exploration of which OTUs are in which groups of samples, and identification of OTUs that explain similarities and differences among samples.

Supplementary References

1. Turnbaugh, P.J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480--484 (2009).
2. Crawford, P.A. *et al.* Regulation of myocardial ketone body metabolism by the gut microbiota during nutrient deprivation. *Proc Natl Acad Sci U S A* **106**, 11276--11281 (2009).
3. Turnbaugh, P.J., Vanessa K. Ridaura, Jeremiah J. Faith, Federico E. Rey, & Gordon, R.K.a.J.I. The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice. *Sci Transl Med* **1** (2009).
4. Quince, C. *et al.* Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6**, 639-641 (2009).
5. Turnbaugh, P.J., Backhed, F., Fulton, L. & Gordon, J.I. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe* **3**, 213-223 (2008).
6. Knight, R. *et al.* PyCogent: a toolkit for making sense from sequence. *Genome Biol* **8**, R171 (2007).
7. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403--410 (1990).

8. Schloss, P.D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**, 7537-7541 (2009).
9. Schloss, P.D. & Handelsman, J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71**, 1501-1506 (2005).
10. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658--1659 (2006).
11. Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5** (2004).
12. Thompson, J.D., Higgins, D.G. & Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673--4680 (1994).
13. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**, 511--518 (2005).
14. Morgenstern, B., Frech, K., Dress, A. & Werner, T. DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics* **14**, 290-294 (1998).
15. Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**, 5261--5267 (2007).
16. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).