# TD Génomique environnementale

Étude d'un consortium microbien

**Version 1.2**

5 et 14 mars 2018

Eléonore FROUIN
eleonore.frouin@mio.osupytheas.fr

# Table des matières

# 0 Rappel sur le système Unix

Présentation des commandes Unix essentielles pour naviguer sur un terminal. L'ouverture du terminal se fait via le gestionnaire d'application ou grâce au raccourci **Ctrl+Alt+T**.

## 0.1 Accéder à une machine distante via ssh

```
$ ssh -X  user@pedaserv2.luminy.univ-amu.fr
```

Pour établir la première connexion ssh, tapez **yes** puis entrez votre mot de passe AMU. L'invite de commande ressemble alors à :

```
user@pedaserv2:~$
```

Pour se déconnecter et revenir sur la machine locale :

```
$ logout
```

## 0.2 Lister, créer et éditer des fichiers

**Liste :** La commande ci-dessous permet de lister les fichiers et les sous répertoires présents dans le répertoire courant. À quoi sert l'option **-l** ?

```
$  ls -l
```

**Arborescence :** Pour naviguer dans l'arborescence des dossiers via un terminal, on utilise la commande **cd** suivi du chemin du dossier cible. Par exemple, pour se placer dans le répertoire **Documents** on utilise :

```
$  cd Documents/
```

Pour remonter vers un dossier parent (dans notre cas, revenir dans le répertoire personnel) :

```
$  cd ../
```

Enfin pour rapidement se placer à la racine de son répertoire personnel, il existe deux solutions :

```
$  cd ~
$  cd
```

**ASTUCE :** Apprenez à utiliser l'**auto-complétion** avec la touche **Tab** pour compléter le chemin et les noms de fichiers existants sans avoir besoin de les écrire en entier.

**Création d'un répertoire** La création du nouveau dossier s'effectue le répertoire courant.

```
$  cd Documents
$  mkdir Unix/
$  mkdir Unix/subdir
```

**Créez un dossier nommé DIR1 dans le répertoire Unix et déplacez-vous dans ce dossier.**

**Édition de fichiers** À quoi sert la commande ci-dessous ?

```
$  touch fichier.txt
```

L'édition de fichier peut être réalisée dans le terminal via des éditeurs de texte tels que vim, emac, nano ..

```
$  nano fichier.txt
```

**Écrivez deux lignes de texte puis quittez l'éditeur en sauvegardant votre fichier.**

Pour connaître le contenu d'un fichier, on utilise la commande **less** ou encore la commande **more**.

```
$  less fichier.txt
```

Taper **q** pour quitter

## 0.3 Copier, déplacer et renommer des fichiers ou dossiers

**Copier** La commande **cp** permet de copier un fichier vers un autre fichier ou dossier.

```
$  cp fichier.txt fichier_copie1.txt
$  cp fichier.txt ../subdir/fichier_copie2.txt
```

La copie d'un dossier nécessite l'ajout de l'option **-r**.

```
$  cd ~/Documents/Unix/
$  cp -r DIR1/ DIR2/
```

**Déplacer** La commande **mv** permet de déplacer un ou plusieurs fichiers et/ou dossiers. L'argument final désigne la destination et est obligatoirement un répertoire.

```
$  cd ~/Documents/Unix/DIR1
$  mv fichier.txt fichier_copie1.txt ../subdir/
```

**Renommer** La commande pour renommer un fichier ou un dossier est également **mv**.

```
$  cd ~/Documents/Unix/DIR2
$  mv fichier.txt Nouveau.txt
```

**Déterminez les actions des commandes ci-dessous (renomme et/ou déplace). On supposera que tous les dossiers en majuscule existent, mis à part le dossier NEW_DIR.**

```
mv test.txt  data
mv test.txt  DATA/
mv test.txt  ../../seq.fasta ../../TEST/
mv seq.fasta ../../TEST/seq.fasta
mv DATA ../../NEW_DIR
```

## 0.4 Supprimer des fichiers ou dossiers

La commande **rm** permet de supprimer des fichiers. Elle doit être complétée par l'option -r pour la suppression des dossiers.

```
$  cd ~/Documents/Unix/
$  rm  DIR2/Nouveau.txt
$  rm -r DIR1
```

Attention : La commande **rm** détruit directement les fichiers, sans passage habituel par la "corbeille".

## 0.5 Quelques dernières commandes : echo, grep

La commande **echo** permet d'afficher une chaîne de caractères sur le terminal (sortie standard).

```
$  echo "j'adore Unix"
$  echo "j'adore Unix"> sortie.txt
```

**Que se passe-t-il ? Que contient le fichier *sortie.txt* ?**

```
$  echo "je ne sais pas"> sortie.txt
```

**Vérifiez à nouveau le contenu du fichier *sortie.txt*.**

```
$  echo "si j'adore Unix" >> sortie.txt
```

**Que fait le $>>$ ?**

La commande **grep** recherche dans un ou plusieurs fichiers les lignes contenant un certain motif. L'option **--color** permet de colorer le motif recherché sur la sortie standard du terminal.

```
$  grep  --color 'Unix' sortie.txt
$  grep  --color 'i' sortie.txt
```

**Expliquez ce que renvoie la commande ci-dessous.**

```
$  grep  -c 'i' sortie.txt
```

# 1 Introduction

## 1.1 Contexte de l'étude

L'article de Koenig *et al* (cf ANNEXE) présente une étude sur la succession des consortia bactériens de la flore intestinale de nourrissons au cours de leur deux premières années. Les auteurs utilisent ici deux techniques : le méta-barcoding et la métagénomique, pour caractériser les communautés microbiennes. Leur analyse les a conduit à deux conclusions :

1. Globalement la diversité phylogénétique et la composition des communautés évoluent graduellement avec le temps.
2. Certains groupes taxonomiques majoritaires peuvent en revanche présenter des brusques changements d'abondance en fonction du régime alimentaire et de la santé.

**Méta-barcoding 16S** Le séquençage ADN du gène de l'ARN ribosomique 16S permet de reconstruire l'histoire évolutive des organismes. Ainsi on peut étudier la taxonomie des espèces présentes, réaliser une analyse de la diversité, et également quantifier les proportions des taxons suivant différentes conditions.

**Métagénomique** Il s'agit du séquençage direct de l'ADN présent dans un échantillon. Dans le cadre de l'article d'étude, il s'agit de matière fécale. La métagénomique consiste à identifier les gènes présents dans la communauté microbienne, et à assigner des fonctions aux gènes identifiés. Cette technique donne un aperçu du potentiel fonctionnel d'un environnement.

## 1.2 Les données

Les deux types de données, utilisée dans l'article de Koenig *et al*, sont à disposition pour aborder l'analyse :

— Des données d'amplicons du gène de l'ARNr 16S (55 échantillons), disponibles sur MG-RAST
`http://metagenomics.anl.gov/?page=MetagenomeProject&project=65`

— Des données de métagénomes (12 échantillons) disponibles sur EBI Metagenomics `https://www.ebi.ac.uk/metagenomics/projects/SRP002437`

## 1.3 Fichiers de séquences biologiques

**Le format FASTA** Format de fichier texte utilisé pour stocker des séquences nucléiques ou protéiques. Une entrée d'un fichier FASTA est constituée de **deux lignes**. La première décrit la séquence en commençant par le signe `">"` suivi de l'identifiant de la séquence. La deuxième ligne contient la séquence en elle-même.

```
> SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
```

L'extension d'un fichier FASTA est conventionnellement *.fasta* ou *.fa*. Il est également possible de trouver les extensions *.fna* et *.faa* pour distinguer les nucléotides des acides aminés.
Le format FASTA se prête facilement à la manipulation et à la lecture des séquences via des outils de traitement de texte et de langages script tels que Perl.

**Le format FASTQ** Format de fichier texte permettant de stocker à la fois des séquences nucléiques et les scores de qualité associés, établis lors du séquençage. Une entrée d'un fichier FASTQ est constituée de **quatre lignes**. La première commence par un caractère `"@"` suivi de l'identifiant de la séquence. La deuxième ligne contient la séquence nucléique. La troisième commence par un caractère "+", parfois suivi d'une description de la séquence. La dernière ligne

contient les scores de qualité associés à chacune des bases de la séquence de la ligne 2. Le score de qualité est codé par un caractère ASCII.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

L'extension d'un fichier FASTQ est conventionnellement *.fastq* ou *.fq*.
Grâce au score qualité associé à chacune des bases, il est possible de filtrer les séquences, en ne conservant que celles dont la qualité est supérieure à un seuil donné.

**Manipulation des fichiers FASTA**

ATTENTION : Les fichiers FASTA peuvent être très volumineux, il est donc déconseillé de les ouvrir dans un éditeur de texte (tel que Gedit). Il est préférable d'utiliser les commandes **less** ou **more** pour les afficher dans un terminal.

**Trouvez la commande Unix qui permet de compter le nombre de séquences dans un fichier fasta. Testez-la sur le fichier *test.fasta*.**

## 1.4 Liste des outils

Il existe de très nombreux outils bioinformatiques pour réaliser chacune des étapes de l'analyse génomique et il n'est pas toujours facile de s'y retrouver. Le site `omictools.com/` recense les principaux. Ces outils, logiciels, peuvent soit être installés en local, soit lancés depuis des interfaces Web (les calculs sont réalisés sur des serveurs distants). Les interfaces Web sont souvent limitées en terme de taille de jeux de données à traiter, mais ne nécessitent aucune installation et ont une prise en main plus aisée pour des non-spécialistes.

## 1.5 Principales étapes du traitement de données métagénomiques

### 1.5.1 Analyse d'amplicons 16S

| | Exemple d'outils bioinfo |
|---|---|
| 1. Filtrage qualité des séquences des amplicons | Trimmomatic |
| 2. Suppression des chimères créées pendant le traitement d'amplification PCR | DECIPHER |
| 3. Alignement des séquences entre elles pour pouvoir les comparer | muscle |
| 4. Regroupement des séquences similaires ou très proches en OTU (Unité Taxonomique Opérationnelle) et sélection pour chaque OTU de la séquence la plus représentative | uclust, usearch61 |
| 5. Assignation et classification des séquences retenues (au niveau du phylum, du genre, de l'espèce ...) | PhymmBL |
| 6. Analyse, comparaison (sortie graphique ..) | phyloseq (R) |

### 1.5.2 Analyse des métagénomes

Les métagénomes contiennent les informations relatives à la composition taxonomique des communautés microbiennes étudiées. Deux méthodes d'analyse sont envisageables ; leurs principales étapes sont présentées ci-dessous.

### A.1. Analyse taxonomique via des gènes spécifiques

1. Filtrage qualité (en fonction du score qualité des fichiers fastq, suppression des réplicats artificiels).
2. Identification de gènes marqueurs de la phylogénie (exemple : gène de l'ARNr 16S, gène mcrA pour les méthanogènes).
3. Affectation de ces gènes à une taxonomie.

**Quel est le principal désavantage de cette méthode ?**

### A.2. Analyse taxonomique basée sur le potentiel protéique

1. Filtrage qualité (fastq.score, suppression des réplicats artificiels)
2. Comparaison de l'ensemble des séquences avec les bases de données protéiques existantes pour déterminer leur taxonomie.

L'analyse des métagénomes permet d'aller bien au-delà de l'analyse taxonomique, en estimant par exemple le potentiel fonctionnel des communautés microbiennes.

### B. Analyse fonctionnelle

1. Filtrage qualité (fastq.score, suppression des réplicats artificiels).
2. Assemblage des séquences génomiques en longs fragments.
3. Détection des cadres ouverts de lectures dans les longs fragments d'ADN.
4. Annotation fonctionnelle des gènes identifiés.

De nombreuses autres analyses sont possibles : reconstruction de voies métaboliques, détermination des relations entre les individus d'une communauté, description de l'histoire évolutive, et même la reconstruction partielle de génomes.

**Quelles sont les techniques d'analyses des métagénomes mises en œuvre dans l'article de Koenig et al ?**

# 2 Tutoriel QIIME

### Arborescence de début de projet

```
QIIME_analysis/
    Analysis/                  # dossier d'analyse
        mapfile.csv            # fichier contenant les métadonnées
        Scripts/               # dossier de scripts de visualisation

    Data16S/                   # dossier contenant les données de l'étude
        AllSeqs.fna            # séquences d'amplicons de 55 échantillons
```

Placez le dossier QIIME_analysis dans votre dossier **Documents** et décompressez l'archive QIIME_analysis.

```
$ tar xvf QIIME_analysis.tar.gz
```

Placez-vous dans le **répertoire Analysis** :
TOUTES LES COMMANDES DOIVENT ÊTRE LANCÉES AU NIVEAU DE CE RÉPERTOIRE.

### Pipeline QIIME

Le pipeline QIIME (acronyme pour Quantitative Insights Into Microbial Ecology) permet de réaliser l'analyse de séquences microbiennes du gène de l'ARNr 16S. Ce pipeline contient une multitude de programmes qui permettent notamment du filtrage qualité, des alignements taxonomiques, reconstructions phylogénétiques, analyses de diversité et visualisations graphiques.

Toutes les commandes QIIME s'effectuent depuis le serveur **pedaserv2**. Pour la connexion à partir d'un terminal, il faut suivre la procédure indiquée dans la partie 0.1. Pour chaque commande QIIME, une aide est accessible en ajoutant l'option ***-h***, notamment pour avoir accès à l'ensemble des options.

```
nom_de_la_commande.py -h
```

## 2.1 Regroupement (=clustering) des séquences en OTUs

Lors de cette première étape toutes les séquences des échantillons sont regroupées en unité taxonomique opérationnelle (OTU, groupe de séquences représentant un certain degré de parenté taxonomique). Ces regroupements sont basés sur la **similarité** entre les séquences.

La ligne de commande ci-dessous permet de regrouper les séquences en OTUs en utilisant un seuil à 97% d'identité entre les séquences pour qu'elles soient regroupées.

```
pick_otus.py -i ../Data16S/AllSeqs.fna -o picked_otus/
```

**Quels sont les fichiers produits et que contiennent-ils ? Combien d'OTUs ont été créées ? Comment faut-il modifier la commande pour un clustering avec 98% d'identité ?**

***Facultatif :*** **Combien de séquences appartiennent au cluster n° 358 ?**

## 2.2 Choix de la séquence la plus représentative par OTU

Chaque OTU est constituée de plusieurs séquences apparentées, l'étape suivante est de sélectionner **une** séquence représentative par cluster. Cette séquence sera utilisée pour l'identification

taxonomique de l'OTU et l'alignement phylogénétique.

```
pick_rep_set.py -i picked_otus/AllSeqs_otus.txt -f ../Data16S/AllSeqs.fna -o Seqs_rep_set.fna
```

**Quel est le format du fichier produit ? Combien contient-il de séquences ?**

### 2.3 Assignation taxonomique des séquences représentatives

Pour affecter une taxonomie aux séquences représentatives choisies, on utilise la méthode uclust consensus taxonomy classifier.

```
assign_taxonomy.py -i Seqs_rep_set.fna -r /usr/local/lib/python2.7/dist-packages/qiime_default_reference/gg_13_8_otus/rep_set/97_otus.fasta
```

**Donner un exemple de séquence ayant une affiliation précise au niveau de l'espèce.**

### 2.4 Alignement des séquences représentatives

Les séquences représentatives des OTUs doivent ensuite être alignées entre elles pour permettre l'analyse phylogénétique. Il est ici réalisé avec PyNAST : en combinant notre jeu de séquences et une base de données de gènes d'ARNr 16S (dans le fichier *85_otus.pynast.fasta*), PyNAST réalise un alignement multiple de l'ensemble de ces séquences.

```
align_seqs.py -i Seqs_rep_set.fna -o Pynast_align -t /usr/local/lib/python2.7/dist-packages/qiime_default_reference/gg_13_8_otus/rep_set_aligned/85_otus.pynast.fasta
```

**À quoi correspondent les '-' dans le fichier *Seqs_rep_set_aligned.fasta* ?**

**Combien de séquences n'ont pas pu être alignées ? Comment pourrait-on diminuer ce nombre ?**

### 2.5 Filtrage de l'alignement

```
filter_alignment.py -i Pynast_align/Seqs_rep_set_aligned.fasta -o filtered_alignment/
```

**À votre avis, à quoi sert cette étape de filtrage d'après la comparaison des fichiers Seqs_rep_set_aligned.fasta et Seqs_rep_set_aligned_pfiltered.fasta ?**

### 2.6 Construction d'un arbre phylogénétique

Dans cette étape on construit un arbre phylogénétique comprenant les séquences représentatives de chaque OTUs. On le visualisera plus tard dans le TP.

```
make_phylogeny.py -i filtered_alignment/Seqs_rep_set_aligned_pfiltered.fasta -o rep_phylo.tre
```

### 2.7 Création d'une table d'OTU

La dernière étape consiste à associer les assignations taxonomiques (cf étape 3) avec les regroupements en OTUs (établis lors de la première étape) pour construire un tableau d'abondance d'OTUs.

```
make_otu_table.py -i picked_otus/AllSeqs_otus.txt -t uclust_assigned_taxonomy/Seqs_rep_set_tax_assignments.txt -o otu_table.biom -m mapfile.csv
```

Le format **.biom** stocke les résultats (qui peuvent être volumineux) de façon compressée. Pour pouvoir lire et analyser ces résultats, il faut convertir le fichier dans un format de fichier texte avec séparateur.

```
biom convert -i otu_table.biom  --to-tsv -o otu_table.tsv --header-key taxonomy
```

**Comment la table des OTUs est elle construite ? (entête de ligne, de colonne ...) ?**

Un résumé de la table d'OTUs peut être obtenu directement (sans convertir le fichier .biom) avec la commande **biom summarize-table**. On peut notamment retrouver le nombre d'OTUs total et par échantillon.

```
biom summarize-table -i otu_table.biom
```

**Enregistrez le résultat de cette commande dans un fichier, que l'on nommera *summary_table_otu.txt* .**

## 2.8 Filtrage de la table d'OTUs

Un post-traitement souvent utile consiste à filtrer certaines OTUs présentes dans la table finale.

**Regardez la première ligne du fichier otu_table.tsv (après l'entête). Qu'en concluez-vous ?**

Par exemple pour ne conserver que les OTUs qui contiennent au minimum 2 séquences, on filtre la table OTUs à l'aide du script filter_otus_from_otu_table.py.

```
filter_otus_from_otu_table.py -i otu_table.biom -n 2 -o otu_table_filtered.biom
```

```
biom convert -i otu_table_filtered.biom  --to-tsv -o otu_table_filtered.tsv
```

```
biom convert -i otu_table_filtered.biom --to-tsv -o otu_table_filtered_taxo.tsv
```

**Combien d'OTUs ont été retenus après filtrage ?**

***Facultatif :* Quelle est l'option qui filtre la table d'OTUs pour garantir que les séquences de chaque OTU proviennent d'au minimum 2 échantillons différents ?**

## 2.9 Analyse et représentation graphique

Il est possible de grouper les OTUs à différents niveaux taxonomiques (phylum, classe, ordre, famille, etc.) grâce au script summarize_taxa_through_plots.py.

```
summarize_taxa_through_plots.py -i otu_table_filtered.biom -o taxa_summary -m mapfile.csv
```

**À quoi correspondent les différentes tables *.txt* créées dans le dossier taxa_summary ?**

# 3 Analyse avec Rstudio

```
biom convert -i otu_table_filtered.biom  - -to-json -o otu_table_filtered.json
```

## 3.1 Script graphes.R

Ce script permet de réaliser des graphes simples à partir des fichiers générés par QIIME.
→ Ouvrir Rstudio
→ Ouvrir le script graphes.R
→ Chaque ligne du script peut être executée grâce au raccourci Ctrl+Entrée
→ Si certaines libraries ne sont pas installées dans votre version de R, il faut les importer via Tools-> Install Packages ...
→ Certaines commandes, signalées par # @@ à compléter/modifier @@, doivent être modifiées avant d'être executées ☺

## 3.2 Script phyloseq.R

Ce script utilise le package phyloseq. Son installation diffère des libraries utilisées jusqu'à présent : il faut copier les lignes ci-dessous dans le terminal R. L'installation est assez lente (environ 15 minutes).

```
source("http ://bioconductor.org/biocLite.R")
biocLite("phyloseq")
```

Le package phyloseq permet le filtrage, le sous-échantilonnage et la comparaison de tables OTUs mais surtout la création de graphiques élaborés.

# 4 Visualisation avec Krona

Krona est un outil de visualisation interactive. Il génère des diagrammes circulaires à plusieurs niveaux offrant une représentation hiérarchique de la classification taxonomique. Cette visualisation permet notamment d'explorer les abondances d'OTUs au sein d'un taxon spécifique en zoomant sur la section d'intérêt. Ces graphiques sont construits à partir d'une table d'OTUs.

```
biom convert -i otu_table_filtered.biom --to-tsv -o otu_table_filtered_taxo2.tsv --header-key  taxonomy
```

```
Scripts/app_otu_to_krona.pl -i otu_table_filtered_taxo2.tsv -o krona_visualization.html
```

La commande ci-dessus génère un fichier html qui peut être ouvert avec un navigateur web, tel que Firefox.

```
firefox krona_visualization.html
```

**Quel pourcentage de *Firmicutes* détecte-t-on dans le microbiote aux jours 4, 58 et 172 ?**

5 ANNEXE

# Succession of microbial consortia in the developing infant gut microbiome

Jeremy E. Koenig[a], Aymé Spor[a], Nicholas Scalfone[a], Ashwana D. Fricker[a], Jesse Stombaugh[b], Rob Knight[b,c], Largus T. Angenent[d], and Ruth E. Ley[a,1]

[a]Department of Microbiology, Cornell University, Ithaca, NY 14853; [b]Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309; [c]Howard Hughes Medical Institute, University of Colorado, Boulder, CO 80309; and [d]Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY 14850

Edited by Todd R. Klaenhammer, North Carolina State University, Raleigh, NC, and approved June 24, 2010 (received for review March 2, 2010)

**The colonization process of the infant gut microbiome has been called chaotic, but this view could reflect insufficient documentation of the factors affecting the microbiome. We performed a 2.5-y case study of the assembly of the human infant gut microbiome, to relate life events to microbiome composition and function. Sixty fecal samples were collected from a healthy infant along with a diary of diet and health status. Analysis of >300,000 16S rRNA genes indicated that the phylogenetic diversity of the microbiome increased gradually over time and that changes in community composition conformed to a smooth temporal gradient. In contrast, major taxonomic groups showed abrupt shifts in abundance corresponding to changes in diet or health. Community assembly was nonrandom: we observed discrete steps of bacterial succession punctuated by life events. Furthermore, analysis of ≈500,000 DNA metagenomic reads from 12 fecal samples revealed that the earliest microbiome was enriched in genes facilitating lactate utilization, and that functional genes involved in plant polysaccharide metabolism were present before the introduction of solid food, priming the infant gut for an adult diet. However, ingestion of table foods caused a sustained increase in the abundance of Bacteroidetes, elevated fecal short chain fatty acid levels, enrichment of genes associated with carbohydrate utilization, vitamin biosynthesis, and xenobiotic degradation, and a more stable community composition, all of which are characteristic of the adult microbiome. This study revealed that seemingly chaotic shifts in the microbiome are associated with life events; however, additional experiments ought to be conducted to assess how different infants respond to similar life events.**

human gut | metagenomics | microbial diversity | community assembly | short chain fatty acids

The assembly of the human gut microbiota begins during birth with colonization by microbes from the environment. In the first few hours of life, the mother's vaginal and fecal microbiomes are usually the most important source of inoculum (1, 2). During the initial few months of a milk diet, bacteria such as Bifidobacteria, highly adapted to process milk oligosaccharides, can be abundant (3). The introduction of solid foods heralds a shift toward bacterial consortia characteristic of the adult microbiota (4).

Although before weaning, the diet is a relatively constant supply of milk, during this time the microbiome can display large shifts in the abundances of bacterial taxa. For instance, in a time series analysis of 14 infants, Palmer et al. (4) documented fluctuations in the abundances of major bacterial taxonomic groups, and the temporal patterns of variation differed between individuals. Interpersonal variation in gut microbial diversity is greater between infants than between adults, and furthermore, the infant microbiome displays more interpersonal variability in functional gene content than the adult microbiome (5). The large functional and phylogenetic variation observed between infant gut microbiomes may be due to random colonization events, differences in immune responses to the colonizing microbes, changes in host behavior, or other aspects of host lifestyle (4, 6). How each of these factors contributes to shaping the infant microbiome remains unclear.

To investigate how life events impact the developing infant gut microbiome, we performed a case study to monitor the gut microbial composition of one infant over a period of 2.5 y. We analyzed a set of more than 60 fecal samples collected concurrently with detailed information regarding diet, health status, and activities. The infant was a full-term, vaginally delivered healthy male. He was placed in a daycare facility during weekdays starting at 3 mo and then removed from group care at 1 y. His diet regimen consisted of exclusive breast-feeding for the first 134 d of life, supplemented with formula until he was no longer breast-fed at 9 mo. The first solid food introduced to the diet was rice cereal at 4 mo, followed by table foods, and the replacement of formula with cow milk at 1 y. The child suffered from several ear infections for which he was treated with antibiotics, but was otherwise healthy, and he was immunized according to the US Centers for Disease Control and Prevention's recommended schedule.

We profiled the bacterial diversity of the fecal samples with 454-pyrosequencing: First, we generated 318,620 16S rRNA gene sequences (Table S1), which we used to map the dynamics of the developing microbiota onto a timeline of changes in diet and other life events. On the basis of the patterns observed from the 16S rRNA gene analysis, we performed a metagenomic analysis of >500,000 sequences from 12 samples to study in greater detail key transitions in microbial community composition triggered by life events (Table S2). These data were used to address the following questions: How does the diversity of the microbiota relate to the functional gene content of the microbiome over time? How are the communities that constitute the microbiota structured? How do changes in diet and events, such as antibiotic treatment, affect the succession and functions of bacteria consortia? This analysis allowed us to pinpoint specific events (e.g., illness, diet change, and antibiotic treatment) likely to have triggered significant changes in this infant's intestinal microbiota.

## Results

**16S rRNA Gene Analysis Reveals Temporal Patterns of Qualitative Diversity.** For each sample, we measured phylogenetic diversity (PD), the sum of all of the branch lengths in a 16S rRNA gene

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Microbes and Health," held November 2–3, 2009 at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. The complete program and audio files of most presentations are available on the NAS Web site at http://www.nasonline.org/SACKLER_Microbes_and_Health.

Author contributions: J.E.K., L.T.A., and R.E.L. designed research; J.E.K., N.S., A.D.F., L.T.A., and R.E.L. performed research; J.S. and R.K. contributed new reagents/analytic tools; J.E.K., A.S., J.S., R.K., and R.E.L. analyzed data; and J.E.K., A.S., A.D.F., and R.E.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All 16S rRNA gene and metagenomic sequence data are archived in GenBank (accession no. SRA012472).

[1]To whom correspondence should be addressed. E-mail: rel222@cornell.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1000081107/-/DCSupplemental and http://microbe.calsnet.cornell.edu/leylab/fileshare/ITS_PNAS_SI/.

phylogenetic tree: the greater the PD, the greater the diversity represented in the sample (7). As expected, PD increased over time and was positively correlated with age ($R^2$ = 0.5; Fig. 1). The first stool sample produced by the infant (meconium, a tarry substance consisting of the in utero accumulation of gut luminal material) had the lowest PD, and the sample with the highest PD was the mother's sample collected on the same day. There are several time points that deviate from the general trend of increasing PD over time (Fig. 1). Day 85, a time point just before a fever, had a low PD compared with preceding days; day 168, when peas and formula were introduced to the diet, had a relatively high PD compared with the previous sample day; and day 195 also had a high PD; however, this was not associated with any documented changes. Two of the three antibiotic treatments are followed by a decrease in PD relative to previous sample days. Although PD for day 244 is located on the trend line illustrated in Fig. 1, it is lower than previous sample days. The second treatment with amoxicillin, however, does not seem to affect the PD of the infant's microbiome as judged by 16S rRNA sequence analysis of sample day 297; this may be an indication of the adaptive power of the human microbome as it pertains to multiple exposures to the same antibiotic. Consistent with the infant's first amoxicillin treatment, a low PD is observed on days 413, 432, and 441 after the infant's first exposure to the antibiotic cefdinir (a broad-spectrum cephalosporin).

In addition to comparing samples using measures of PD, we performed a principal coordinates analysis (PCoA) of unweighted UniFrac (8) to determine how the diversity among samples changed during the sampling period. This analysis showed that the diversity changed gradually over time (Fig. 2 *A–D*). Fecal samples collected early in the time series harbored microbial communities more similar to one another than to samples collected later on, and vice versa. The samples that deviate from this diversity gradient, days 413, 432, and 441, are the samples noted above with a lower relative PD. Samples associated with the same diet are adjacent in the gradient because they were collected from the same period in the infant's life. For instance, breast-milk, formula, and solid food associated samples form a contiguous pattern in the PCoA plot (Fig. 2 *B–D*).

**Succession of Bacterial Consortia and Patterns of Quantitative Diversity.** The abundance of operational taxonomic units (OTUs) was assessed across all samples, and OTUs were clustered in a heat map according to their cooccurrence (Fig. 3*A*). This clustering analysis revealed a succession of bacterial communities that resolved four discrete phases (steps) initiated by life events (e.g., fever at day 92 separates step 1 from step 2, diet change at day 161 divides steps 2 and 3, and antibiotic treatment and adult diet at day 371 divides steps 3 and 4). A linear discriminant analysis (LDA) was carried out to assess the statistical significance of these four steps: the a posteriori assignment probabilities of the steps indicate whether the fecal samples can be properly assigned to the steps given their community structure. Thus, we assigned the four steps as a priori categories in the LDA, and the resulting posterior probabilities for steps 1–4 were 0.90, 0.64, 0.76, and 0.71, respectively (Table S3). These results indicate that the steps can be differentiated according to the bacterial consortia of their respective fecal samples.

In step 1 (days 3–84; Fig. 3*A*), the gut microbiome comprises a specific suite of Firmicute OTUs. Step 2 is preceded by an increase in the abundance of proteobacterial OTUs (days 92–100), which coincided with fever symptoms. Actinobacterial and proteobacterial OTU abundances increased in step 2, and the suite of Firmicute OTUs observed in step 1 differed from those observed in step 2. The introduction of formula and peas to the infant's diet is associated with an increase in bacteroidetes in step 3 (days 172–297) that continues in step 4 (days 454–838); however, the specific Bacteroidetes OTUs enriched differ between these two steps (Fig. 3 *A* and *B*). The transition phase (days 371–441) from steps 3 to 4 is characterized by a number of environmental changes, including cefdinir treatment for an ear infection, exclusion of breast milk and formula from the diet, and an introduction of cow milk and a full adult diet. Interestingly, the transition phase preceding step 4 comprises OTUs that are typical of those observed during step 1, and therefore appear as outliers; again, these are the same samples that are outliers in the PD and UniFrac patterns (Figs. 1 and 2). Because this is a case study, we cannot attribute any single life event as the definitive pressure leading to the formation of the gut microbiomes defined within step 4. One scenario is that this change in the infant's microbiome may have been induced by a purge in PD as a result of cefdinir treatment. The microbial landscape in the gut could then reform according to substrates that are typical of an adult diet. Regardless, the abundances of bacterial phyla are relatively constant in step 4: this constancy among samples col-

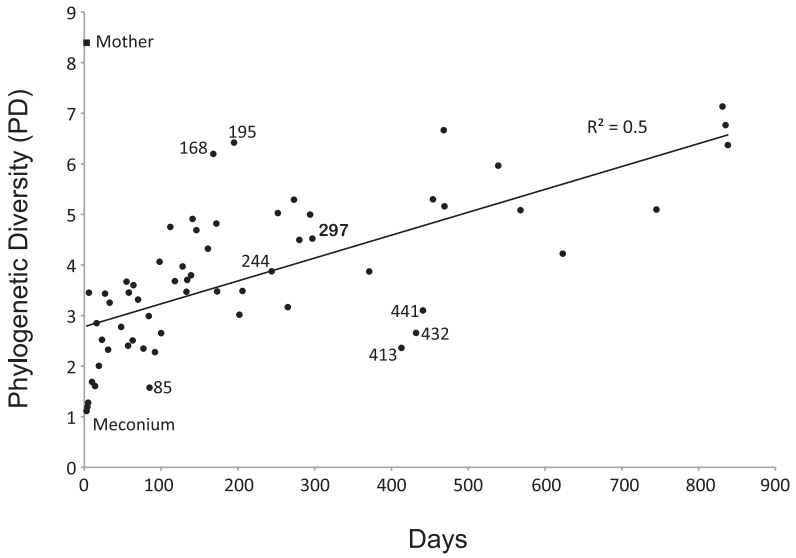


**Fig. 1.** Bacterial PD of the infant gut microbiota over time. PD provides a measure of the diversity within a community based on the extent of the 16S rRNA phylogenetic tree that is represented by that community. Symbols are fecal samples. The mother's fecal sample, collected at day 3, is denoted as a filled square.

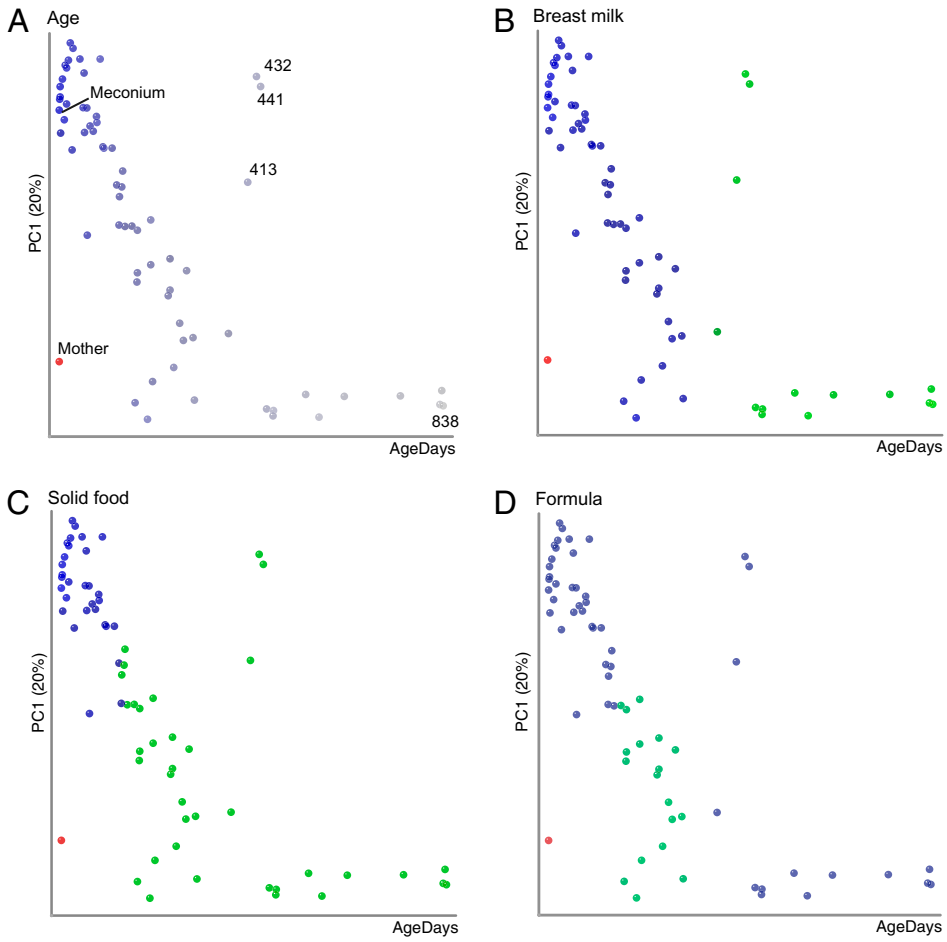


**Fig. 2.** Community composition changes over time conform to a smooth temporal gradient. Time and PC1 from a PCoA of bacterial communities determined from 16S rRNA genes are plotted. (*A*) The color gradient corresponds to time (days): earlier samples are darkest blue, and later samples are paler. The mother's sample, in red, was assigned an age of 3 d and clusters along PC1 with the older samples from the infant. (*B–D*) Same data projection as in *A*; however, days for which breast milk was part of the diet are in blue in *B*, days with solid food (including rice cereal) are green in *C*, and days with formula are green in *D*. Symbols are individual fecal samples; the variance explained by PC1 is indicated on the axis.

lected over more than 400 d is an indication that the infant gut microbiome has reached a stable state.

**Species Cooccurrence and Exclusions.** Because our OTU-based cluster analysis revealed a succession of different microbial consortia over time (Fig. 3*A*), we tested whether the developing infant's gut microbiota was subject to community assembly rules. Specifically, we invoked two measures that assess OTU cooccurrence: the C-score and checkerboard measures. The C-score assesses the tendency for species to exclude one another from a given niche (9), whereas the number of checkerboard pairs corresponds to the number of species pairs that never cooccur (10). To assess the significance of the scores obtained from the dataset, we compared the C-score and checkerboard indices from actual data with scores obtained from 5,000 communities assembled randomly from the same OTU data. The C-score for the real dataset was 38.97, which is significantly greater than the simulated mean C-score of 35.98 obtained from the randomized data ($P < 0.0002$). The checkerboard measure for the microbial communities (2,561.00) was also significantly greater than the randomized mean checkerboard measure (2, 321.03, $P < 0.0002$; Fig. 4 *A* and *B*). Together, these ecological measures indicate that the developing infant gut microbiota is composed of interacting bacterial consortia, not of randomly assembled suites of bacteria.

**Bacterial Load and Diversity in Relation to Short Chain Fatty Acid Concentrations.** To gain insight into how community structure relates to microbial metabolite pools, we checked for relationships between bacterial diversity and short chain fatty acid (SCFA) concentrations in fecal samples (Fig. 5 *A–C*). Specifically, we measured the concentration of acetate, propionate, and butyrate in 56 fecal samples by GC-MS, and bacterial load by quantitative PCR. Overall, levels of acetate were highest and butyrate lowest, and levels of all three SCFAs were highly correlated with each other (Fig. 5*B*). SCFA levels and bacterial load were generally higher after the introduction of solid foods (Figs. S1 and S2). Bacterial diversity was correlated with all three SCFAs: PC1 of the unweighted UniFrac PCoA was negatively correlated with all three SCFAs ($R^2$: 0.3, 0.4, and 0.1 for acetate, propionate, butyrate, respectively, $P < 0.001$). A regularized canonical correlation analysis (RCCA) indicated that Bacteroidetes abundances were positively correlated with all three SCFAs and most strongly with propionate levels (Fig. 5*C* and Fig. S3). Verrucomicrobia were also positively correlated with acetate and propionate levels. In contrast, the abundance of Firmicutes correlated negatively with all three SCFAs and most strongly with propionate. Collectively, these measures suggest that community assembly is nonrandom and likely reflects syntrophic and antagonistic relationships mediated by microbial metabolites.

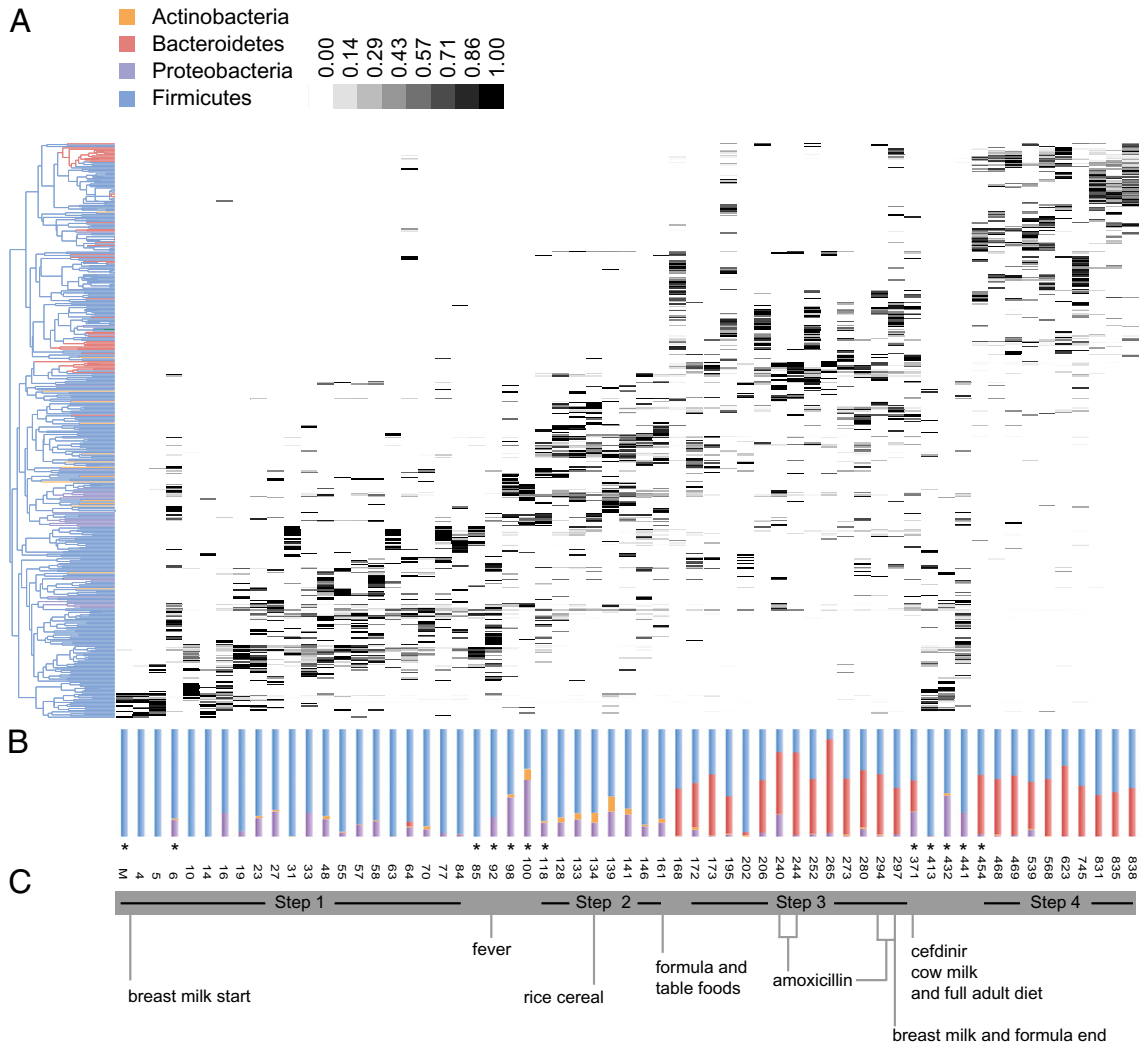


**Fig. 3.** OTU-based community structure and composition in the gut microbiota. (*A*) Each vertical lane corresponds to a sample day, and the gray-scale shaded rectangles represent the abundance of the different OTUs. The dendogram on the left shows how the OTUs are clustered according to cooccurrence, and branches are colored to indicate the taxonomical assignment of the OTUs at the phylum level. Samples selected for metagenomic analyses are indicated with asterixes. (*B*) Relative abundances of the bacterial phyla in each samples. (*C*) Significant events pertaining to changes in the infant's diet are indicated. Steps characterized by specific bacterial consortia supported by linear discriminate analysis are shown.

**PD of Metagenomic Sequences.** Taxonomic assignment was determined using BLASTX (11), and the majority of sequences were bacterial genes. Low levels of fungi and viruses were also detected, and Euryarcheaota (Archaea) were detected in all samples including meconium (<0.01% of sequences). The majority of DNA sequences extracted from fecal samples collected at the beginning of the time series (meconium and day 6) were assigned to the Firmicute phylum (Fig. 6*A*), which is consistent with our PCR-based 16S rRNA gene survey for day 6 (Fig. S4). However, our 16S rRNA gene results for days 92–118, which show an abundance of Firmicute OTUs, are inconsistent with the abundance of actinobacterial genes obtained from these samples, likely reflecting 16S rRNA gene primer bias. Furthermore, the metagenomic analysis recovered fewer proteobacterial genes compared with the 16S rRNA gene-based analysis. Nevertheless, the patterns obtained from these two methods are consistent overall for this time period: the highest levels of actinobacterial and proteobacterial sequences were observed on sample days 92–118 in both analyses (Fig. S4). Interestingly, day 92, which was associated with fever, has the highest viral and fungal levels (Fig. 6*A*). Later in the time series (days 413, 432, and 441 after diet change and cefdinir treatment), the relative decrease in levels of bacteroidetes OTUs observed by 16S rRNA analysis was not observed in our BLASTX taxonomic assignment of metagenomic sequences (Fig. S4).

**Functional Gene Dynamics in the Developing Infant Gut Microbiome.** We used the Meta Genome Rapid Annotation using Subsystem Technology (MG-RAST) (12) to assign gene functions to the 12 metagenomic samples. A summary of these results is represented as normalized heat maps, also generated using MG-RAST (Fig. 6*B*). According to relative abundances of subsystems, samples clustered into three main groups that reflect the time period of sample collection (Fig. 6*B*). We ran bootstrapping and resampling analyses to identify genes that were enriched in samples relative to an average representation of genes across the 12 samples (Table S4). Analysis of the meconium sample (day 3) revealed an enrichment of carbohydrate-metabolizing genes involved in lactose/galactose and sucrose uptake and utilization, genes involved in antibiotic resistance (e.g., ABC transporters), and virulence genes (e.g., multidrug resistance efflux genes, adhesion proteins, and pathenogenicity islands). Day 6 also had many of the same enriched gene functions as the meconium samples, in addition to gene functions associated with cell mem-

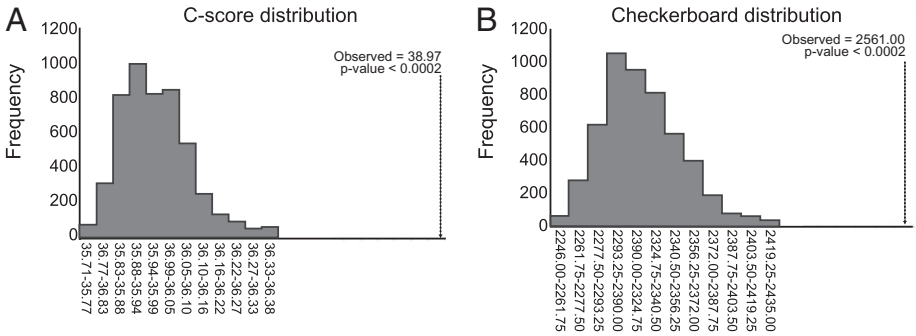


**Fig. 4.** Community assembly is nonrandom. (*A*) C-score distributions for observed and randomized OTU occurrence in each sample. (*B*) Checkerboard indices for observed and randomized OTU occurrence. Values for the observed distributions are indicated with arrows.

brane and cell wall components (Table S4). Furthermore, on day 6, genes associated with vitamin biosynthesis (e.g., vitamin B12, folate) were already present in the infant microbiome. By day 85, carbohydrate-using genes for amylose, arabinose, and maltose degradation, and virulence genes such as type III and IV secretion systems are enriched (Table S4). At day 92, when fever occurred, enriched eukaryotic rRNA modification genes likely reflect higher relative levels of fungi (Table S4). Carbohydrate-using genes enriched on day 92 include rhamnose, fructooligosaccahride and raffinose-utilization pathways, and xylose-degradation genes. Enrichment of sialic acid metabolism genes (day 85) and β-glucoronide utilization genes (day 100) may indicate that the microbiota is capable of using or mimicking host glycans early in life. Furthermore, in days 98, 100, and 118, before the introduction of the first solid food, additional genes for the utilization of plant-derived glycans, such as xylitol, are present (Table S4).

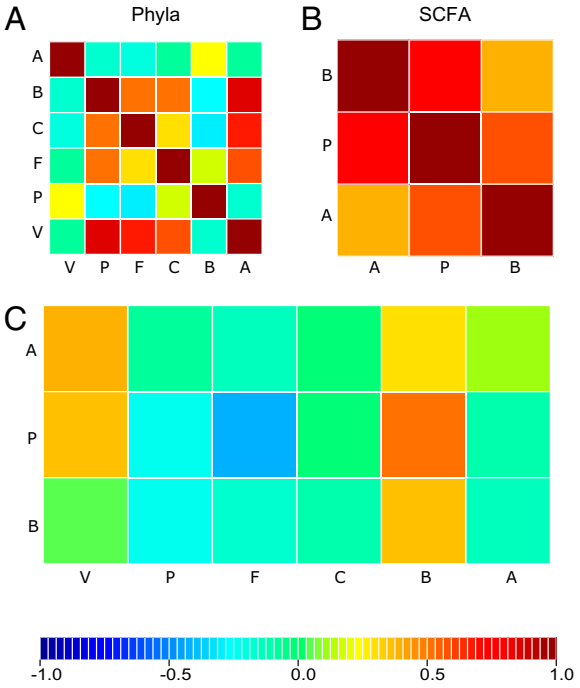


**Fig. 5.** Relationships between phyla abundances and levels of SCFAs in feces. (*A–C*) Correlation matrices. (*A*) Phyla (V, Verrucomicrobia; P, Proteobacteria; F, Firmicutes; C, Cyanobacteria; B, Bacteroidetes; A, Actinobacteria). (*B*) SCFAs (A, acetate; P, propionate; B, butyrate). (*C*) Cross-correlation between phylum abundances and SCFA concentrations. The color scale indicates that negative correlation values are in blue tones, whereas positive correlation values are red.

At the later time points (days 371, 431, 441, and 454) a complement of genes associated with the adult microbiome's core metabolic functions, namely polysaccharide breakdown, vitamin biosynthesis, and xenobiotic degradation, are evident. For instance, on day 371, genes for the utilization of maltose, maltodextrin, xylose, and mannose, which are polysaccharide breakdown products, are enriched. In addition, vitamin and cofactor biosynthesis genes including vitamin B6, thiamin, and flavodoxin are enriched on these sample days. Finally, genes reflecting the diversity of substrates in an adult diet were recovered; for example, genes for cinnamic acid degradation (day 432), benzoate catabolism (day 441), and additional enzymes involved in the anaerobic degradation of aromatic compounds (day 454) are present.

**Relating Function to Phylogeny in the Infant Gut Microbiome.** We used RCCA to compare samples according to their gene content (Fig. S3 *A–D*). One step in RCCA is to correlate the abundances of phyla (from the phylogenetic assignment of genes) across samples; this revealed that the abundances of genes assigned to the Firmicute, Bacteroidetes, and Euryarcheal phyla were positively correlated. In addition, the actinobacterial and proteobacterial gene content of samples was positively correlated (Fig. S3). RCCA resolved clusters of samples and indicated which functional genes were driving the clustering (Fig. S3 and Tables S5 and S6). Meconium, day 6, and day 85 form a cluster because they are enriched in genes taxonomically assigned to the Firmicutes, and their functions include Gram-positive cell wall components and central carbohydrate and organic acid metabolism (Fig. S3 *A* and *B* and Table S6). The sample from day 92, associated with fever, is clearly separated from the other samples in the analysis because it is enriched with genes assigned to the fungal phylum (Fig. S3 *C* and *D* and Table S6). The following days (98, 100, and 118) also separate from other metagenomic samples and are characterized by genes encoding ABC transporters and assigned phylogenetically to the Actinobacteria and Proteobacteria, (Fig. S3*D* and Table S6). Interestingly, the abundance of actinobacterial and proteobacterial genes is strongly negatively correlated to the abundance of Firmicute genes. Days 371, 432, 441, and 454 are clustered because of their Bacteroidetes gene content, and this pattern is driven by an enrichment in genes related to carbohydrate fermentation, Gram-negative cell wall, and capsule formation (Fig. S3 *A* and *B* and Table S6).

## Discussion

An essential goal of the human microbiome project is to understand the assembly and community composition of the microbiota, not only to gain a better understanding of our own biology, but also because the microbiome is implicated in human health (13). Gut microbiotas can contribute to excess host adiposity (14–16), protect against the development of type 1 diabetes (17), and induce colitis (18) and metabolic syndrome (19).

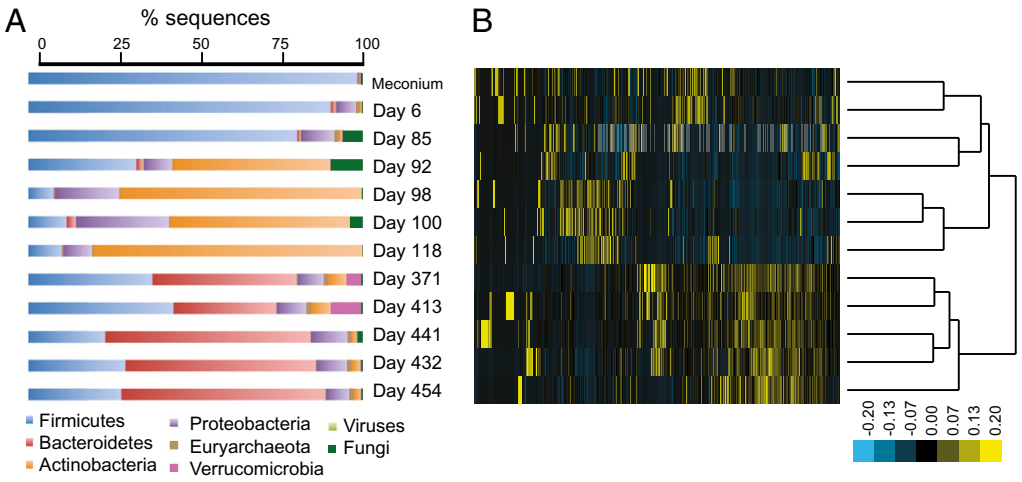


**Fig. 6.** Metagenomic analysis of DNA sequences extracted from infant fecal DNA. (*A*) Taxonomic assignment of metagenomic sequences. (*B*) Heat map and hierarchical clustering of samples based on MG-RAST subsystem gene content.

Thus, the microbiota has been suggested as a target for therapeutic intervention for several chronic diseases (13, 20–22). Adult microbiotas are thought to be relatively stable over time (14, 23, 24); this stability imparts resilience to disturbance, ensuring continued gut function. In a disease context, however, such stability and resilience could be detrimental if the gut community is pathogenic. Understanding the succession of bacterial consortia in the human gut during childhood may help in the development of strategies to guide the formation of health-promoting microbiotas that could then be maintained throughout the life of the host.

Our study of the gut microbiome of one infant followed over a 2.5-y period allowed an in-depth look into the dynamics of a developing intestinal ecosystem in relation to known disturbances. We observed a gradual increase in diversity over time, related to a gradual change in community diversity. Superimposed on these patterns of gradual change are the effects of life events, such as drastic diet changes or antibiotic treatments, which result in large shifts in the relative abundances of taxonomic groups. The qualitative measures of diversity, such as PD and UniFrac, responded to time, but the quantitative measures, such as the specific abundances of OTUs assembled into consortia of interacting species, responded to life events. Additional studies considering multiple subjects will assess whether infant microbiomes respond consistently to the same life events.

Our metagenomic analyses provided additional insight into the dynamics of the developing microbiome. For instance, the infant suffered a fever at day 92, during the exclusively breast-milk–fed period, which is followed by a shift in the abundances of a specific suite of OTUs. Fungal and viral genes were enriched at that time, suggesting a transient imbalance in the microbiota that might have been directly related to the fever. Another noteworthy observation was that genes facilitating the breakdown of plant-derived polysaccharides were present during this period, despite an exclusive breast-milk diet. This second observation is consistent with other metagenomic analyses of infant gut microbiomes, which reported microbial enzymes that degrade nondigestible polysaccharides of plant origin (2, 5). Together these studies suggest that the infant microbiome is metabolically ready for receiving simple plant-derived foods, such as rice cereal. This may explain why the introduction of rice cereal did not result in detectable changes in the 16S rRNA gene profiles in this intant's gut microbiome.

The introduction of peas and formula, followed by other table foods, may have been the cause of a codominance of the Bacteroidetes and Firmicutes and enrichment in functional genes characteristic of the adult gut microbiome. In addition to carbohydrate-using genes used for the breakdown of plant polysaccharides, functional genes present in the weaned infant microbiome included those involved in the breakdown of xenobiotic compounds and in vitamin biosynthesis. The abundances of bacterial phyla were relatively constant after weaning, indicating that the infant gut microbiome has reached a stable state. Together these results suggest that the 2.5-y-old human gut microbiome has many of the functional attributes of the adult microbiome.

The fine-scale temporal sampling allowed us to test whether the gut microbial community was subject to ecological assembly rules over time. The C-score and checkerboard analyses, which test for species cooccurrence and exclusion, strongly support a nonrandom pattern of community assembly. The human gut microbiota is known to be composed of syntrophic partners (25), as well as competing members (26, 27). Such ecological interactions likely underlie the nonrandom associations of species constituting the microbiota.

The introduction of table foods was followed by a large shift in phyla abundances within the infant's microbiome, in addition to increased bacterial loads and SCFA levels. Although specific members of the Firmicute phylum, such as *Roseburia* spp., are known to produce butyrate and respond to carbohydrate levels in the diet (28), this analysis did not detect positive relationships between Firmicute OTUs and SCFA levels, perhaps because a wide variety of gut bacteria can produce these metabolites. However, our 16S rRNA gene analysis showed a dramatic and sustained increase in the abundance of Bacteroidetes immediately after the introduction of peas and other table foods to the diet. The Bacteroidetes are specialized in the breakdown of complex plant polysaccharides (29); the introduction of plant-derived carbohydrates into the diet could have boosted populations of Bacteroidetes, which is consistent with mouse microbiome studies (30). The metabolic activities of these Bacteroidetes may have either directly or indirectly increased production of SCFAs. Consistent with these observations, low levels of Bacteroidetes in the gut are correlated with obesity, which itself may result from a diet low in plant-derived polysaccharides (23, 31). Thus, together these results further support the notion that a diet high in plant material promotes a microbial community structure and metabolite production that is beneficial to the human host.

This study revealed the power of sampling a microbiome over time to gain insight into the events that can alter its phylogenetic and functional composition. Our results complement those of Palmer et al. (4), who documented large compositional shifts in the abundances of major bacterial taxa over time in 14 babies,

which they postulated could be a reflection of life events. We also observed large shifts in the abundances of major groups; interestingly, these shifts are associated with life events, such as illnesses, dietary changes, and antibiotic treatment, suggesting that differences in the colonization patterns of multiple babies would most likely reflect differences in their daily lives. Indeed, future temporal human microbiome studies should be performed in parallel to assess whether individual microbiomes respond differently to the same disturbances.

## Methods

**Samples and DNA Extraction.** This study was approved by the Internal Review Board of Washington University in St. Louis (protocol no. 09-0039), and samples were transferred to Cornell University under protocol no. 0910000952. Fecal samples were collected from a full-term, healthy infant during diaper changes. The birth was vaginal, no antibiotics were administered to the mother or the baby at birth, and the mother was antibiotic-free for the duration of the pregnancy. Samples were immediately frozen upon collection at −20 °C, then transferred to the laboratory and maintained at −80 °C until processing. Frozen samples were ground under liquid $N_2$, then a subsample of ≈100 mg was used for whole-community DNA extraction. A 100-mg aliquot of each homogenized sample was suspended while frozen in a solution containing 500 mL of DNA extraction buffer [200 mM Tris (pH 8.0), 200 mM NaCl, and 20 mM EDTA], 210 mL of 20% SDS, 500 mL of a mixture of phenol/chloroform/isoamyl alcohol (25:24:1), and 500 mL of a slurry of 0.1-mm-diameter zirconia/silica beads (BioSpec Products). Microbial cells were then lysed by mechanical disruption with a bead beater (BioSpec Products) set on high for 2 min (22 °C), followed by extraction with phenol/chloroform/isoamyl alcohol and precipitation with isopropanol. The quantity and quality of purified DNA was assessed using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen) and a plate reader.

**Sample Preparation for 454 Pyrosequencing of 16SrRNA Genes.** 16S rRNA genes were amplified from each sample using a composite forward primer and a reverse primer containing a unique 12-base barcode, which was used to tag PCR products from respective samples (31). We used the forward primer 5′-*GCCTTGCCAGCCCGCTCAGTC***AGAGTTTGATCCTGGCTCAG**-3′: the italicized sequence is 454 Life Sciences primer B, and the bold sequence is the broadly conserved bacterial primer 27F. The reverse primer used was 5′-*GCCTCCCTCGCGCCATCAGNNNNNNNNNNNNCA*-**TGCTGCCTCCCGTAGGAGT**-3′: the italicized sequence is 454 Life Sciences' primer A, and the bold sequence is the broad-range bacterial primer 338R. NNNNNNNNNNNN designates the unique 12-base barcode used to tag each PCR product (31, 32), with "CA" inserted as a linker between the barcode and rRNA primer. PCR reactions consisted of HotMaster PCR mix (Eppendorf), 200 μM of each primer, and 10–100 ng template, and reaction conditions were 2 min at 95 °C, followed by 30 cycles of 20 s at 95 °C, 20 s at 52 °C, and 60 s at 65 °C on an Eppendorf thermocycler. Three independent PCRs were performed for each sample, combined and purified with Ampure magnetic purification beads (Agencourt), and products visualized by gel electrophoresis. No-template extraction controls were analyzed for lack of visible PCR products. Products were quantified using Quant-iT PicoGreen dsDNA assay as described above. A master DNA pool was generated from the purified products in equimolar ratios to a final concentration of 21.5 ng $mL^{-1}$. The pooled products were sequenced using a Roche 454 FLX pyrosequencer at the Cornell University LIfe Sciences Core Laboratories Center.

**16S rRNA Gene Sequence Analysis.** Sequences generated from pyrosequencing barcoded 16S rRNA gene PCR amplicons (average length 237 nt; Table S1) were analyzed using default settings in the open source software package Quantitative Insights Into Microbial Ecology (QIIME; http://qiime.sourceforge.net). 16S rRNA gene sequences were assigned to OTUs using the QIIME implementation of cd-hit (33) and a threshold of 97% pairwise identity. OTUs were classified taxonomically using the Ribosomal Database Project (RDP) classifier 2.0 (34). A single representative from each OTU was aligned using PyNast (35) to build the phylogenetic tree used to for measuring the PD of samples (7) and unweighted UniFrac (36).

***Cooccurrence analysis.*** The C-score and checkerboard indices (9) were determined using a null hypothesis of random community assembly, whereby 5,000 matrices were randomly generated from the 16S rRNA gene 0.97 OTU data with EcoSim Version 7.0. C-score and checkerboard distributions and *P* values were determined from the simulations using EcoSim's default settings.

***Clustering analysis.*** Rarified (randomly subsampled to normalize sequence counts) OTUs, with an abundance greater than 5% and present in two or more samples, were hierarchically clustered using Kendall's τ similarity metric. The Self Organizing Map was generated using 20,000 iterations, also using the Kendall's τ similarity metric, in the freeware Cluster 3.0 (http://www.falw.vu/~huik/cluster.htm). Heat map graphics were generated using JavaTreeView (37). An LDA was carried out in R for studying multivariate clustering of fecal samples according to their associated microbiotas (abundances of different RDP-assigned classes).

**Metagenomic Analysis of the Infant Gut Microbiome.** A metagenomic analysis was used to assess the diversity of microbial genes within the infant gut microbiome at different sample days. We studied three time periods: the early infant gut microbial communities (the meconium at day 3, and day 6), days associated with fever (days 85–118), and one time range associated with cefdinir treatment and diet change (days 371–454). Twelve whole-community fecal DNA samples were barcoded, pooled, and shotgun sequenced using the Roche-454 Titanium pyrosequencer. After filtering low-quality reads, we obtained a total of 482,919 sequences (Table S2).

Metagenomic sequences were trimmed using the CLC Genomic Work Bench 3.0. The minimum allowable sequences length was 100 bp, quality score limit was 0.05, only two ambiguous nucleotides were permitted per sequence, and a hit limit of moderate was used to identify and remove vector sequences. The 454 replicate filter software (38) was used to remove sequences that were artificially replicated during the sequencing protocol. Filtered nucleotide metagenomic sequences were compared with the September 27th, 2009 version of the National Center for Biotechnology Information nonredundant database (nr) using BLASTX (11), and results were visualized in MEGAN (39) to determine the taxonomic distribution of genes in each library (i.e., the best BLASTX result using a maximum e-score of $10^{-5}$ was used as an approximation for the taxonomic origin of a given sequence). Metagenomic sequences were functionally annotated using MG-RAST (http://metagenomics.nmpdr.org), built as a modified version of the RAST server (12). Normalized heat maps were also generated using MG-RAST, and the different gene pool arrays were hierarchically using Cluster 3.0. An RCCA was performed to highlight correlations between the phylum abundance matrix (X of order $n \times p$) and the gene functions matrix (Y of order $n \times q$) retrieved from metagenomics as well as bacterial phyla and SCFAs using the R software CCA package (40). Regularization parameters $\lambda_1$ and $\lambda_2$ were chosen to maximize the leave-one-out cross-validation score (41).

**Quantitative PCR Analysis.** Real-time PCR amplification and detection were performed using an ABI 7300 Real Time PCR System (Applied Biosystems). We used the Power Sybr Green PCR Master Mix (Applied Biosystems), including 0.2 μM of 16S rRNA primers 8F (5′ AGAGTTTGATCCTGGCTCAG) and 338R (5′ CTGCTGCCTCCCGTAGGAGT). Cycling conditions included an initial incubation of 50 °C for 2 min, denaturing at 95 °C for 10 min, then 40 cycles of 95 °C for 15 s, 60 °C for 1 min, and a dissociation curve step of 95 °C for 15 s, 60 °C for 30 s, and 95 °C for 15 s.

**SCFA Analysis.** For each sample, 200 mg of frozen feces was vortexed for 1 min in 1% HCl. Isotope-labeled SCFAs were added in a final concentration of 5 mM [1-$^{13}$C] acetate, 1 mM [$^2H_5$] propionate (Cambridge Isotopes), and 1 mM [$^2H_5$] propionate (Sigma Aldrich). Homogenized samples were centrifuged at 2,350 × *g* for 30 s. Supernatant was acidified to pH 0 with HCl. Each sample was partitioned into four aliquots and extracted at 4 °C with an equal volume of diethyl ether. Samples were incubated with 1-tertbutyl-dimethylsilyl-imidazole (Sigma Aldrich) at 60 °C for 30 min before GC-MS analysis (Agilent 5975C Series; Agilent Technologies).

**ACKNOWLEDGMENTS.** We thank Jeffrey Gordon for his support and Jeffrey Werner for comments on the manuscript. This research was supported by National Human Genome Research Institute grants (to R.K.) and an Arnold and Mabel Beckman Foundation Young Investigator award (to R.E.L.).

1. Gueimonde M, et al. (2006) Effect of maternal consumption of lactobacillus GG on transfer and establishment of fecal bifidobacterial microbiota in neonates. *J Ped Gastroenterol Nutr* 42:166–170.
2. Vaishampayan PA, et al. (2010) Comparative metagenomics and population dynamics of the gut microbiota in mother and infant. *Genome Biol Evol* 2010:53–66.
3. Sela DA, et al. (2008) The genome sequence of *Bifidobacterium longum* subsp. *infantis* reveals adaptations for milk utilization within the infant microbiome. *Proc Natl Acad Sci USA* 105:18964–18969.
4. Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO (2007) Development of the human infant intestinal microbiota. *PLoS Biol* 5:e177, 10.1371/journal.pbio.0050177.

5. Kurokawa K, et al. (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* 14:169–181.
6. Dethlefsen L, Eckburg PB, Bik EM, Relman DA (2006) Assembly of the human intestinal microbiota. *Trends Ecol Evol* 21:517–523.
7. Faith DP (1992) Conservation evaluation and phylogenetic diversity. *Biol Conserv* 61: 1–10.
8. Lozupone C, Knight R (2005) UniFrac: A new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71:8228–8235.
9. Stone L, Roberts A (1990) The checkerboard score and species distributions. *Oecologia* 85:74–79.
10. Diamond JM (1975) Assembly of species community. *Ecology and Evolution of Communities*, eds Cody ML, Diamond JM (Harvard Univ Press, Cambridge, MA), pp 342–444.
11. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
12. Meyer F, et al. (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9: 386, 10.1186/1471-2105-9-386.
13. Turnbaugh PJ, et al. (2007) The human microbiome project. *Nature* 449:804–810.
14. Turnbaugh PJ, et al. (2009) A core gut microbiome in obese and lean twins. *Nature* 457:480–484.
15. Ley RE, et al. (2005) Obesity alters gut microbial ecology. *Proc Natl Acad Sci USA* 102: 11070–11075.
16. Turnbaugh PJ, et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444:1027–1031.
17. Wen L, et al. (2008) Innate immunity and intestinal microbiota in the development of Type 1 diabetes. *Nature* 455:1109–1113.
18. Garrett WS, et al. (2007) Communicable ulcerative colitis induced by T-bet deficiency in the innate immune system. *Cell* 131:33–45.
19. Vijay-Kumar M, et al. (2010) Altered gut microbiota in toll-like receptor-5 deficient mice results in metabolic syndrome. *Science* 328:228–231.
20. Zaneveld J, et al. (2008) Host-bacterial coevolution and the search for new drug targets. *Curr Opin Chem Biol* 12:109–114.
21. Jia W, Li H, Zhao L, Nicholson JK (2008) Gut microbiota: A potential new territory for drug targeting. *Nat Rev Drug Discov* 7:123–129.
22. Rautava S, Kalliomäki M, Isolauri E (2005) New therapeutic strategy for combating the increasing burden of allergic disease: Probiotics—a Nutrition, Allergy, Mucosal Immunology and Intestinal Microbiota (NAMI) Research Group report. *J Allergy Clin Immunol* 116:31–37.
23. Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: Human gut microbes associated with obesity. *Nature* 444:1022–1023.
24. Dethlefsen L, Huse S, Sogin ML, Relman DA (2008) The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol* 6:e280, 10.1371/journal.pbio.0060280.
25. Gibson GR, Macfarlane GT, Cummings JH (1993) Sulphate reducing bacteria and hydrogen metabolism in the human large intestine. *Gut* 34:437–439.
26. Duncan SH, et al. (2003) Effects of alternative dietary substrates on competition between human colonic bacteria in an anaerobic fermentor system. *Appl Environ Microbiol* 69:1136–1142.
27. Flint HJ, Duncan SH, Scott KP, Louis P (2007) Interactions and competition within the microbial community of the human colon: Links between diet and health. *Environ Microbiol* 9:1101–1111.
28. Duncan SH, et al. (2007) Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. *Appl Environ Microbiol* 73:1073–1078.
29. Xu J, et al. (2003) A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science* 299:2074–2076.
30. Turnbaugh PJ, et al. (2009) The effect of diet on the human gut microbiome: A metagenomic analysis in humanized gnotobiotic mice. *Science Transl Med* 1:16ra14.
31. Costello EK, et al. (2009) Bacterial community variation in human body habitats across space and time. *Science* 326:1694–1697.
32. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* 5: 235–237.
33. Li W, Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.
34. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73:5261–5267.
35. Caporaso JG, et al. (2010) PyNAST: A flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26:266–267.
36. Hamady M, Lozupone C, Knight R (2010) Fast UniFrac: Facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J* 4:17–27.
37. Saldanha AJ (2004) Java Treeview—extensible visualization of microarray data. *Bioinformatics* 20:3246–3248.
38. Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J* 3:1314–1317.
39. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17:377–386.
40. Gonzlez I, Déjean S, Martin PGP, Baccini A (2008) CCA: An R package to extend canonical correlation analysis. *J Stat Softw* 23:1–14.
41. Leurgans S, Moyeed R, Silverman B (1993) Canonical correlation analysis when the data are curves. *J R Stat Soc Ser B* 55:725–740.